Université de Lille

École doctorale Biologie-Santé
Laboratoire LilNCog (UMR-S1172) et CRIStAL (UMR 9189)
Équipe PSY et 3D-SAM

Deep Learning for Simulation in Healthcare. Application to Affective Computing and Surgical Data Science

Apprentissage Profond pour la Simulation en Santé. Application à l'Informatique Affective et à la Science des Données Chirurgicales

Par Kevin Feghoul

Thèse de doctorat spécialité Mathématiques et leurs interactions

Devant un jury composé de :

Denis Hamad Professeur, Université du Littoral Côte d'Opale, France	Président du jury
Astrid Herrero Professeur, Université de Montpellier, France	Rapporteure
Stefano Berretti Associate Professor, University of Florence, Italy	Examinateur
Deise Santana Maia Maître de Conférence, Université de Lille, France	Examinatrice
Ali Amad Professeur, Université de Lille, France	Directeur de thèse
Mohamed Daoudi Professeur, IMT Nord Europe, France	Co-Directeur de thèse

Abstract

In this thesis, we address various tasks within the fields of affective computing and surgical data science that have the potential to enhance medical simulation. Specifically, we focus on four key challenges: stress detection, emotion recognition, surgical skill assessment, and surgical gesture recognition. Simulation has become a crucial component of medical training, offering students the opportunity to gain experience and refine their skills in a safe, controlled environment. However, despite significant advancements, simulation-based training still faces important challenges that limit its full potential. Some of these challenges include ensuring realistic scenarios, addressing individual variations in learners' emotional responses, and, for certain types of simulations, such as surgical simulation, providing objective assessments. Integrating the monitoring of medical students' cognitive states, stress levels and emotional states, along with incorporating tools that provide objective and personalized feedback, especially for surgical simulations, could help address these limitations. In recent years, deep learning has revolutionized the way we solve complex problems across various disciplines, leading to significant advancements in affective computing and surgical data science. However, several domain-specific challenges remain. In affective computing, automatically recognizing stress and emotions is challenging due to difficulties in defining these states and the variability in their expression across individuals. Furthermore, the multimodal nature of stress and emotion expression introduces another layer of complexity, as effectively integrating diverse data sources remains a significant challenge. In surgical data science, the variability in surgical techniques across practitioners, the dynamic nature of surgical environments, and the challenge of effectively integrating multiple modalities highlight ongoing challenges in surgical skill assessment and gesture recognition. The first part of this thesis introduces a novel Transformer-based multimodal framework for stress detection that leverages multiple fusion techniques. This framework integrates physiological signals from two sensors, with each sensor's data treated as a distinct modality. For emotion recognition, we propose a novel multimodal approach that employs a Graph Convolutional Network (GCN) to effectively fuse intermediate representations from multiple modalities, extracted using unimodal Transformer encoders. In the second part of this thesis, we introduce a new deep learning framework that combines a GCN with a Transformer encoder for surgical skill assessment, leveraging sequences of hand skeleton data. We evaluate our approach using two surgical simulation tasks that we have collected. Additionally, we propose a novel Transformer-based multimodal framework for surgical gesture recognition that incorporates an iterative multimodal refinement module to enhance the fusion of complementary information from different modalities. To address existing dataset limitations in surgical gesture recognition, we collected two new datasets specifically designed for this task, on which we conducted unimodal and multimodal benchmarks for the first dataset and unimodal benchmarks for the second.

Résumé

Dans cette thèse, nous abordons diverses tâches dans les domaines de l'informatique affective et de la science des données chirurgicales qui ont le potentiel d'améliorer la simulation médicale. Plus précisément, nous nous concentrons sur quatre défis clés : la détection du stress, la reconnaissance des émotions, l'évaluation des compétences chirurgicales et la reconnaissance des gestes chirurgicaux. La simulation est devenue un élément important de la formation médicale, offrant aux étudiants la possibilité d'acquérir de l'expérience et de perfectionner leurs compétences dans un environnement sûr et contrôlé. Cependant, malgré des avancées significatives, la formation basée sur la simulation fait encore face à d'importants défis qui limitent son plein potentiel. Parmi ces défis figurent la garantie de scénarios réalistes, la prise en compte des variations individuelles dans les réponses émotionnelles des apprenants, et, pour certains types de simulations, comme les simulations chirurgicales, l'évaluation objective des performances. Intégrer le suivi des états cognitifs, des niveaux de stress et des états émotionnels des étudiants en médecine, ainsi que l'incorporation d'outils fournissant des retours objectifs et personnalisés, en particulier pour les simulations chirurgicales, pourrait aider à pallier ces limitations. Ces dernières années, l'apprentissage profond a révolutionné notre façon de résoudre des problèmes complexes dans diverses disciplines, entraînant des avancées significatives en informatique affective et en science des données chirurgicales. Cependant, plusieurs défis spécifiques à ces domaines subsistent. En informatique affective, la reconnaissance automatique du stress et des émotions est difficile en raison des problèmes de définition de ces états et de la variabilité de leur expression chez les individus. De plus, la nature multimodale de l'expression du stress et des émotions ajoute une couche de complexité supplémentaire, car l'intégration efficace de sources de données diverses demeure un défi majeur. En science des données chirurgicales, la variabilité des techniques chirurgicales entre les praticiens, la nature dynamique des environnements chirurgicaux, et l'intégration de plusieurs modalités soulignent les difficultés pour l'évaluation automatique des compétences chirurgicales et la reconnaissance des gestes. La première partie de cette thèse propose un nouveau cadre de fusion multimodale basé sur le Transformer pour la détection du stress, en exploitant plusieurs techniques de fusion. Ce cadre intègre des signaux physiologiques provenant de deux capteurs, chaque capteur étant traité comme une modalité distincte. Pour la reconnaissance des émotions, nous proposons une approche multimodale innovante utilisant un réseau de neurones convolutifs sur graphes (GCN) pour fusionner efficacement les représentations intermédiaires de plusieurs modalités, extraites à l'aide de Transformer encoders unimodaux. Dans la deuxième partie de cette thèse, nous introduisons un nouveau cadre d'apprentissage profond qui combine un GCN avec un Transformer encoder pour l'évaluation des compétences chirurgicales, en exploitant des séquences de données de squelettes de mains. Nous évaluons notre approche en utilisant des données issues de deux tâches de simulation

chirurgicale que nous avons collectées. Nous proposons également un nouveau cadre multimodal basé sur le Transformer pour la reconnaissance des gestes chirurgicaux, intégrant un module itératif de raffinement multimodal afin d'améliorer la fusion des informations complémentaires entre différentes modalités. Pour pallier les limitations des ensembles de données existants en reconnaissance des gestes chirurgicaux, nous avons collecté deux nouveaux ensembles de données spécifiquement conçus pour cette tâche, sur lesquels nous avons effectué des benchmarks unimodaux et multimodaux pour le premier ensemble de données et des benchmarks unimodaux pour le second.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Thesis Objectives

The integration of simulation-based training in medical education has been driven by the need to enhance the learning experience of medical students. Medical simulations offer a controlled and immersive setting where students can refine their skills, make decisions under pressure, and learn from their mistakes without real-world consequences. Yet, as these simulations become more sophisticated, several challenges remain unaddressed.

The growing complexity of simulation-based medical training brings to light several key challenges that limit its full potential. These challenges include ensuring realistic training scenarios that mirror the pressures of real-life clinical settings, effectively monitoring cognitive load and psychological states during training, and providing objective and personalized feedback to medical students, particularly in simulations such as surgical training. Addressing these issues is crucial to advancing the effectiveness of simulation-based training.

Particularly in the context of surgical simulations, the main challenges revolve around the lack of objectivity in performance assessment and the difficulty in providing precise, actionable feedback. Current evaluation methods often rely on subjective ratings from seniors surgeons, which come with several limitations, such as being time-consuming, having inconsistent evaluation standards, introducing bias, and potentially intimidating medical students. In addition, while existing surgical simulation tools can provide procedure-level assessments that offer a broad overview of a trainee's capabilities, they often lack the ability to deliver gesture-level evaluations that could provide more detailed insights. This level of granularity is important for effective skill development in surgical training, as it enables targeted feedback that can identifie precise areas where medical students face difficulties, leading to more focused skill improvement and accelerated learning.

In response to these challenges, there is growing interest in leveraging advanced technologies such as deep learning to enhance both educational outcomes and the personalization of simulation-based training. The objective of this thesis is to develop novel deep learning models targeting two important domains in healthcare that could significantly benefit medical simulations: affective computing and surgical data science.

Affective Computing: One focus of my research is the improvement of existing stress detection and emotion recognition state-of-the-art methods. Building on the foundational work of Yujin Wu [1], a former PhD student from our research lab, whose approach was integrated into simulation-based software, my role was to explore more advanced machine learning models and improve the performances of her proposed methodologies. Although the models were validated using public datasets not directly linked to medical simulation, the findings demonstrate the relevance of these methods and underscore their potential applicability in medical training environments. The developed models could be integrated into simulation environments to monitor trainees' emotional and stress levels in real time, allowing for the dynamic adjustment of training scenarios based on each student's psychological state. This capability would support personalized training experiences, enhancing the overall learning process and better preparing students for high-pressure, real-world medical situations.

It is important to note that while this thesis broadly addresses emotion recognition, leveraging a public dataset that includes a range of affective states, including pain, the work presented in Chapter 5 is specifically articulated around pain detection. This focus is primarily driven by the existing research landscape, where this dataset is commonly used for evaluating and comparing novel methods for pain detection. The choice of focusing on pain for classification is intended purely for benchmarking purposes, ensuring meaningful comparisons with state-of-the-art approaches. However, any other affective state from the dataset could have been selected without significantly altering the technical approach or the applicability of the models. The underlying techniques and models developed are versatile and applicable to the wider array of emotional states present in the dataset, including the non-pain class, which comprises multiple emotions.

Surgical Data Science: My research in this domain focuses on surgical skill assessment and surgical gesture recognition. The accurate evaluation of a trainee's proficiency during simulation exercises, as well as the ability to automatically recognize surgical gestures during surgical procedures, is important for offering actionable feedback. By integrating these capabilities into simulation-based training, learners can receive more targeted recommendations to refine their techniques and improve their overall competency in surgical practices. This approach promises to elevate the quality of surgical education by providing data-driven insights that align with best practices and real-world expectations.

Overall, the models developed in this thesis aim to provide key insights, whether about the stress and emotional states of medical students during any kind of medical simulation scenario or in more specialized simulations like surgical-based ones by providing key performance metrics, to enhance the simulation experience.

1.2 Thesis Challenges

The development of robust deep learning models regarding the tasks previously discussed is not without challenges. In the following, we will outline the key challenges specific to each domain:

1.2.1 Affective Computing

Automatic stress detection and emotion recognition face several challenges due to the complexity of human emotions and the limitations of current technology. Below, we present the main challenges associated with these tasks:

Multimodal Fusion: Affective computing tasks benefit greatly from multimodal learning because human emotions are inherently complex and often expressed through a combination of cues across multiple modality, such as facial expressions, vocal tones, body language, and physiological signals. Despite this advantages, multimodal fusion presents several key challenges. First, aligning data from different modalities with varying temporal resolutions and structures is complex, as it requires precise synchronization to ensure corresponding

features are correctly associated. Furthermore, while some modalities provide complementary insights, others may introduce noise or irrelevant information, making it essential to balance feature correlation and minimize redundancy to enhance performance. Managing missing or noisy data is another critical issue, especially in real-world scenarios where one modality may be unreliable or unavailable. Additionally, selecting the appropriate level of fusion, whether at the input level (early fusion) or decision level (late fusion), is crucial, as each approach involves trade-offs in terms of model complexity and performance. To fully exploit the strengths of different modalities, it is often essential to develop more advanced fusion techniques that go beyond traditional methods.

Emotion Complexity and Labeling: Emotions are dynamic, multifaceted, and often ambiguous, making them difficult to define and categorize accurately. Traditional categorical emotion labels capture only a limited range of the full emotional spectrum. More sophisticated approaches, such as continuous dimensions (e.g., valence-arousal) can provide a more comprehensive representation but complicate the labeling process. Furthermore, annotating emotional data remains a challenge, largely due to the subjectivity involved, different annotators may interpret the same expression differently, leading to inconsistencies. Additionally, emotions change over time and across contexts, requiring time-sensitive annotations to capture transitions between emotional states accurately.

Behavior Variability: Behavioral variation poses significant challenges for automatic stress detection and emotion recognition systems due to the inherent variability in how emotions and stress are expressed both within and between individuals. On an intra-subject level, emotional responses and stress behaviors can differ significantly for the same person depending on the context, time, and situation. On an inter-subject level, people express emotions and stress in highly diverse ways, often influenced by personality, culture, and individual coping mechanisms. While one person might react to stress with visible signs like facial expressions or gestures, another might internalize stress with few external indicators, making it harder for the model to generalize across different individuals. This variation is especially critical when aiming for personalized and accurate predictions.

Dataset Constraints: Existing datasets in affective computing are often limited in size, lack diversity, biased toward specific demographics, and collected in controlled environments. Emotional expressions in these datasets are frequently exaggerated or artificial, diverging from the natural behavior observed in real-world settings. This reduces the effectiveness of models when applied outside controlled conditions. Addressing these issues requires large-scale, ecologically valid datasets that capture authentic emotional expressions across a broad demographic spectrum.

1.2.2 Surgical Data Science

Surgical skill assessment and surgical gesture recognition faces several challenges due to the complexity of surgical procedures and the diverse sources of data involved. Below, we present the main challenges associated with this domain:

Multimodal Fusion: Similar to affective computing tasks, surgical-based recognition models can benefit significantly from multimodal learning, as different data sources, such as

video recordings and kinematics data, capture different aspects of the surgical procedure. However, integrating these heterogeneous data types presents significant challenges. Surgical data modalities differ in temporal resolution, data structure, and representation format (e.g., video sequences, kinematic signals), making synchronization and alignment complex. Feature correlation and redundancy must be carefully managed to avoid noise, while still capturing complementary information from different data sources. To address these challenges, advanced fusion techniques are often required to effectively integrate surgical data modalities.

Variability in Surgical Techniques: Surgical skill assessment and gesture recognition face the challenge of significant variability among surgeons in how they perform the same procedure. This variability can be due to individual differences in training, experience, and preferred techniques. Models must be robust enough to account for these variations while still identifying the key factors that differentiate skill levels. Developing models that generalize across different surgeons, procedures, and skill levels is a non-trivial task.

Complex Surgical Environments: Surgical environments are dynamic and often involve numerous challenges such as motion artifacts, occlusions, and the presence of multiple instruments. Real-time assessment requires models to deal with cluttered scenes where instruments might overlap or where the surgeon's hands and tools are partially obscured.

Annotation Complexity in Surgical Skill Assessment: Surgical skill assessment faces several challenges that complicate the accurate evaluation of a surgeon's performance. One significant issue is the lack of objective and standardized metrics for skill evaluation. While traditional assessments often rely on subjective evaluations by senior surgeons, these can be inconsistent and biased, varying based on the evaluator's experience and personal criteria. This subjectivity creates a need for more objective, data-driven measures that can provide consistent and reproducible assessments.

Annotation Complexity in Surgical Gesture Recognition: Annotating datasets for surgical gesture recognition is challenging due to the complexity of gestures, the need for expert annotators, and the time-consuming, costly nature of the process. Annotators must label the beginning and end of each gesture in the videos, which often involve subtle, continuous movements that are difficult to segment, leading to high inter-annotator variability. Ambiguous gesture boundaries and visual noise, and occlusions further complicate the task, affecting annotation precision and model accuracy.

1.3 Thesis Contributions

Building on the challenges outlined in the preceding section, this thesis makes several key contributions to the fields of affective computing and surgical data science, with the aim of advancing simulation-based medical education. The contributions are focused on the development of novel deep learning models that address key challenges in stress detection, emotion recognition, surgical skill assessment, and surgical gesture recognition. The primary contributions of this research are outlined below:

Multimodal Fusion for Affective Computing: We introduces two key innovations in the domain of multimodal learning for the tasks of stress detection and emotion recognition:

- For stress detection, we propose a novel multimodal fusion framework based on the Transformer [2] model that leverages multiple fusion techniques to integrate physiological signals from two sensors attached to the human body, treating each sensor's data as a distinct modality. We outperformed the litterature by a large margin.

- For emotion recognition, we propose MMGT, a novel deep learning multimodal fusion framework that leverages Graph Convolutional Networks (GCNs) [3] to model interactions across different levels of modality-specific representations derived from multiple data sources, including facial landmarks, facial action units, and physiological data. These representations are first extracted using unimodal Transformer encoders, and the relationships between them are captured using a graph-based structure. Our proposed approach achieves state-of-the-art performance.

Surgical Skill Assessment: We propose a novel deep learning framework, STGFormer, for evaluating surgical skills using hand skeleton data sequences. To the best of our knowledge, this is the first framework to leverage hand skeleton sequences for the automatic assessment of surgical expertise. STGFormer integrates a GCN to learn spatio-temporal representations of hand movements by exploiting the natural graph structure of the hand skeleton. Additionally, it incorporates a Transformer encoder to capture long-range dependencies within these representations. Our framework achieves state-of-the-art performance on two simulation-based surgical tasks, effectively distinguishing between the performances of attending surgeons and surgical residents in surgical simulated procedures.

Surgical Skill Assessment Datasets: We present two novel datasets specifically designed for the assessment of surgical skills in two simulated tasks: circular cutting and needle passing. These datasets provide high-quality video data for evaluating surgical skills, supporting the development and validation of methods in this domain. The first dataset includes circular cutting tasks performed on the VirtaMed medical simulator by 4 attending surgeons and 12 surgical residents. The second dataset comprises needle passing exercises conducted by 7 attending surgeons and 22 surgical residents using the same simulator.

Multimodal Fusion Framework for Surgical Gesture Recognition: We propose MGRFormer, a novel attention-based multimodal framework specifically designed for surgical gesture recognition. Our approach introduces an iterative multimodal refinement module that enhances the fusion of complementary information from both kinematic and video modalities at the refinement level. Unlike previous work on surgical gesture recognition, which lacks a refinement mechanism, MGRFormer is, to the best of our knowledge, the first to explore multimodal fusion at the refinement stage, facilitating a more context-aware and temporally coherent understanding of surgical gestures. The multimodal refinement module iteratively improves predictions, allowing the model to correct errors and better capture subtle gesture nuances. By refining predictions in a multimodal context, the model effectively learns cross-modal dependencies and resolves ambiguities that might arise when using either modality in isolation. MGRFormer significantly outperforms classical multimodal fusion techniques and state-of-the-art methods by a substantial margin.

Surgical Gesture Recognition Datasets: We present two novel surgical simulation datasets specifically designed for surgical gesture recognition, addressing the limitations of existing datasets. The first dataset consists of multiple executions of the peg transfer task performed by attending surgeons and surgical residents. This dataset includes videos of the procedures alongside corresponding surgical tool trajectories, tracked using a state-of-the-art deep learning-based object detection model. The second dataset focuses on a suturing task performed multiple times by attending surgeons and medical students. It includes first-person video recordings that capture the entire field of vision during the suturing procedure. Subsequently, we conducted both unimodal and multimodal surgical gesture recognition benchmarks for the peg transfer dataset and performed unimodal surgical gesture recognition benchmarking for the suturing dataset.

1.4 Thesis Outline

The manuscript is organized as follows. Chapter 2 covers the necessary background on deep learning techniques and methodologies that are essential for understanding the novel approaches introduced in this thesis. Chapter 3 provides the theoretical foundation on human emotions, presenting multiple emotion models, the different modalities through which emotions are expressed, and a literature review of methods for emotion recognition using these modalities. Chapters 4 and 5 introduce our proposed multimodal deep learning frameworks for stress detection and emotion recognition, respectively. Chapter 6 presents a novel deep learning framework for surgical skill assessment along with two newly collected datasets specifically designed for evaluating this framework. Chapter 7 proposes a Transformer-based multimodal framework for surgical gesture recognition, followed by Chapter 8, which presents two newly collected datasets specifically for the aforementioned task. Finally, Chapter 9 summarizes the work presented in this thesis and discusses potential directions for future research. The manuscript is divided into two parts: Part I: Affective Computing (Chapters 3, 4, and 5) and Part II: Surgical Data Science (Chapters 6, 7, and 8).

1.5 Publications

The research conducted for this thesis has led to several peer-reviewed publications, which serve as the basis for this manuscript. These publications are listed below:

Kevin Feghoul, Deise Santana Maia, Mohamed Daoudi, Ali Amad. MMGT: Multimodal Graph-based Transformer for Pain Detection. 31st European Signal Processing Conference (EUSIPCO 2023), pp. 556-600.

Kevin Feghoul, Deise Santana Maia, Mehdi El Amrani, Mohamed Daoudi, Ali Amad. Spatial-Temporal Graph Transformer for Surgical Skill Assessment in Simulation Sessions. 26th Iberoamerican Congress on Pattern Recognition (CIARP 2023), pp. 287-297.

Kevin Feghoul, Deise Santana Maia, Mohamed Daoudi, Ali Amad. Transformer multimodal pour la détection du stress. 22ème édition de la conférence COmpression et REprésentation des Signaux Audiovisuels (CORESA 2023).

Kevin Feghoul, Deise Santana Maia, Mehdi El Amrani, Mohamed Daoudi, Ali Amad. MGR-Former: A Multimodal Transformer Approach for Surgical Gesture Recognition. The 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024).

Chapter 2

Deep Learning Background

Contents

This chapter provides essential background on key deep learning techniques that are fundamental to the approach proposed in the thesis.

We begin in Section 2.1 by introducing the Transformer architecture, which has become the backbone of many state-of-the-art models for sequential data processing across various domains. Next, in Section 2.2, we present the Graph Neural Networks architecture, which have gained prominence for their ability to model relational data and learn over graph structures. Finally, in Section 2.3, we provide an overview of the field of Multimodal Machine Learning. As modern intelligent applications increasingly involve processing and integrating information from multiple sources, such as text, images, and signal, understanding how to effectively combine and reason across multiple modalities is critical.

2.1 The Transformer

2.1.1 Introduction

The rapid advancement in natural language processing (NLP) over the past decade has been largely driven by the development of sophisticated neural network architectures. Traditional sequence models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), have demonstrated their capability to handle sequential data by maintaining a hidden state (memory) that captures the information from previous time steps. The hidden state h_t at time step t is computed as:

$$h_t = f(h_{t-1}, x_t)$$

where f is a nonlinear function, typically implemented as a neural network, and x_t is the input at time step t.

Despite their effectiveness, RNNs suffer from several limitations:

- **Vanishing and Exploding Gradients:** During training, gradients can become very small (vanish) or very large (explode), making it difficult to learn long-range dependencies.
- **Sequential Processing:** RNNs process input sequences one element at a time, which hinders parallelization and leads to inefficient training, especially for long sequences.
- **Long-Term Dependencies:** Capturing dependencies over long sequences is challenging, as the influence of earlier inputs diminishes over time.

The introduction of the Transformer architecture in the paper "Attention is All You Need" [2] marked a paradigm shift in sequence modeling. Unlike RNNs and LSTMs, the Transformer model relies entirely on a attention mechanism to capture global dependencies between input and output sequences. This architecture not only enhances the ability to

model long-range dependencies but also allows for significantly more parallelization during training and testing, leading to drastic improvements in both performance and efficiency.

The Transformer model has since become foundational in a wide range of NLP tasks, including machine translation, text generation, and sentiment analysis. Its architecture has inspired subsequent developments such as BERT [4] and the GPT series—GPT [5], GPT-2 [6], and GPT-3 [7]. Among these advancements, ChatGPT, built on the GPT-3 architecture, exemplifies the practical application of these models in generating human-like conversational responses, showcasing the significant potential of Transformers in creating interactive AI systems.

The impact of Transformer-based models extends beyond NLP. In computer vision, Vision Transformers (ViTs) [8] have achieved state-of-the-art performance on image classification tasks. Moreover, Transformers have demonstrated state-of-the-art performance across various other domains, including speech recognition [9], computational biology [10], reinforcement learning [11], time-series forecasting [12], and multimodal learning [13], among others.

2.1.2 Model Architecture

The Transformer architecture consists of an encoder and a decoder, each composed of a stack of identical layers. The encoder processes the input sequence and produces a continuous representation, while the decoder generates the output sequence from this representation. Figure 2.1 illustrates the overall structure of the Transformer architecture. In the following sections, we will delve deeper into each component of the Transformer.

Encoder-Decoder Blocks

Below is a breakdown of the components of the encoder and decoder:

- **Encoder:** The encoder is composed of a stack of N identical layers, each consisting of two main sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Furthermore, residual connections are applied around each of these sub-layers, followed by layer normalization. Both sub-layers and the embedding layers produce outputs of dimension d_{model}.

- **Decoder:** Similarly, a stack of N identical layers, but each layer has an additional multi-head attention mechanism that performs attention over the encoder's output. Like the encoder, residual connections and layer normalization are used around each sub-layer.

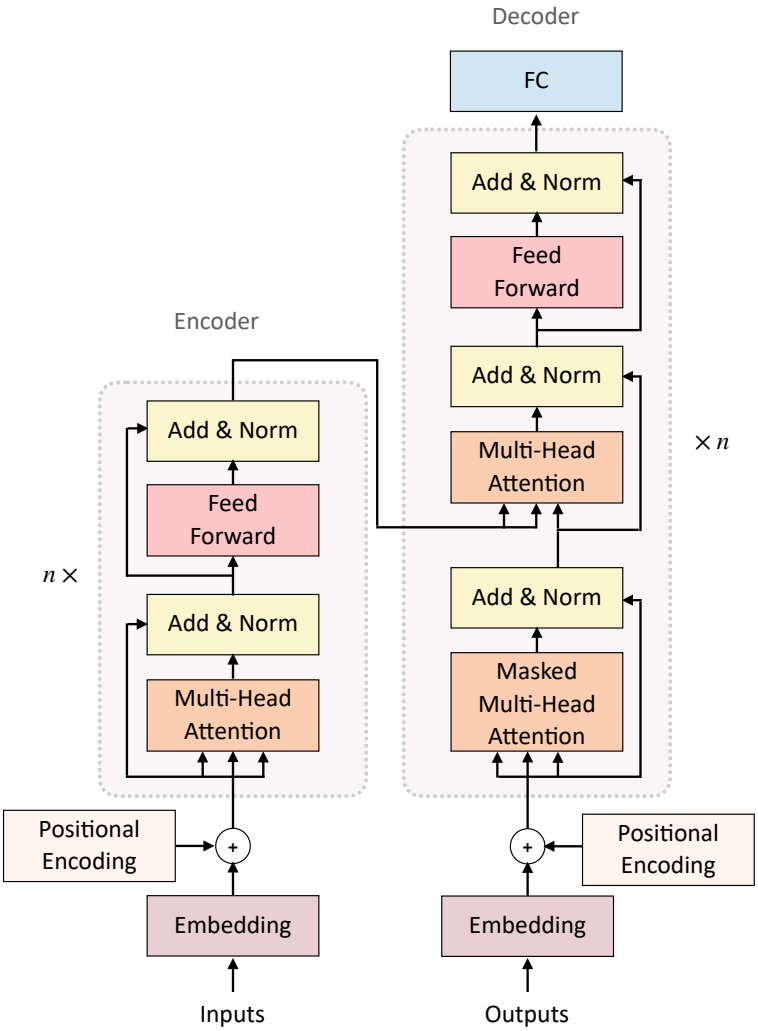


Figure 2.1: The Transformer architecture [2].

Attention

The attention mechanism was first introduced by Bahdanau et al. [14] in the context of machine translation. Their approach allowed the model to selectively focus on specific parts of the input sequence while generating each element of the output sequence. The attention mechanism was designed to enhance the performance of deep learning models by enabling them to identify and concentrate on the most relevant portions of the input data. This innovation has significantly boosted performance across various tasks, including machine translation, text summarization, and question answering. Prior to the introduction of attention mechanisms, models like RNNs and LSTMs faced challenges such as vanishing gradients and difficulty in capturing long-range dependencies in sequences. The attention mechanism addresses these issues by enabling models to dynamically prioritize different parts of the input sequence as they produce each output element.

The core idea behind attention is to compute a weighted sum of values (V), where the weights are determined by a compatibility function between a query (Q) and a set of keys

(K). This allows the model to dynamically attend to different parts of the input sequence, depending on the context, and thus capture more complex dependencies. In the context of the Transformer model, attention mechanisms play a central role in processing input data, making it possible to handle long-range dependencies efficiently.

Scaled Dot-Product Attention: The Scaled Dot-Product Attention is the fundamental building block of the attention mechanism in Transformers. It operates on queries and keys of dimension d_k, and values of dimension d_v, which are all vectors derived from the input sequences. The process can be broken down into the following steps:

1. **Dot Product**: Compute the dot product between the query and all keys to obtain a set of scores. These scores indicate how much focus the model should place on each key-value pair.

$$\text{Scores} = QK^T$$

2. **Scaling**: Scale the scores by the square root of the dimension of the keys (d_k) to prevent the dot products from growing too large, which can lead to very small gradients. This scaling factor stabilizes the training process.

$$\text{Scaled Scores} = \frac{QK^T}{\sqrt{d_k}}$$

3. **Softmax**: Apply the softmax function to the scaled scores to obtain the attention weights. The softmax function normalizes the scores into probabilities, which sum to one.

$$\text{Attention Weights} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

4. **Weighted Sum**: Compute the weighted sum of the values, using the attention weights. This produces the output of the attention mechanism, which is a combination of the input values, weighted by their importance.

$$\text{Output} = \text{Attention Weights} \cdot V$$

In summary, the Scaled Dot-Product Attention mechanism allows the model to focus on different parts of the input sequence dynamically, improving its ability to capture dependencies and relationships within the data.

Multi-Head Attention: While the Scaled Dot-Product Attention provides a mechanism for focusing on relevant parts of the input, it might still be limited in its ability to capture diverse patterns of relationships within the data. Multihead Attention extends this capability

by allowing the model to jointly attend to information from different representation subspaces at different positions. The Multihead Attention mechanism involves the following steps:

1. **Linear Projections**: The input queries, keys, and values are linearly projected h times with different, learned linear projections matrices. This results in h different sets of queries, keys, and values:

$$Q_i = XW_i^Q, \quad K_i = XW_i^K, \quad V_i = XW_i^V \quad \text{for } i = 1, \ldots, h$$

where X is the input, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ are the projection matrices for the i-th head.

2. **Parallel Attention**: Each set of projected queries, keys, and values is then passed through the Scaled Dot-Product Attention mechanism in parallel, resulting in h different outputs:

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$$

3. **Concatenation**: The outputs from the h attention heads are concatenated and linearly projected with a learned projection matrix to produce the final output:

$$\text{MultiheadOutput} = W^O \left[\text{head}_1; \text{head}_2; \ldots; \text{head}_h\right]$$

where $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ is the outpout projection matrix.

By using multiple attention heads, the model can capture a richer set of dependencies from different subspaces of the input. Each head can focus on different aspects of the input data, providing a more comprehensive representation.

In the original paper, the authors set $d_{model} = 512$ and $h = 8$ attention heads, with the dimensions of the key and value vectors fixed at $d_k = d_v = d_{model}/h = 64$.

Position-Wise Feed-Forward Networks

Each encoder and decoder layer also includes a position-wise feed-forward network, which consists of two linear layer with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

This feed-forward network is applied independently to each position in the sequence, providing non-linear transformations that enhance the model's capacity to capture complex patterns.

The purpose of the second linear layer is to map the activated output from a higher-dimensional space back to the original input dimension (or some other output dimension).

In most Transformer architectures, the dimensionality of the input and output of the FFN remains the same. The intermediate hidden layer (activated by ReLU) typically has a higher dimensionality, allowing for a richer representation before reducing back to the original dimensionality.

Layer Normalization and Residual Connections

To stabilize and accelerate training, each sub-layer in the encoder is followed by layer normalization and employs residual connections. The output of each sub-layer can be described as:

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

where x is the input to the sub-layer, and $\text{Sublayer}(x)$ represents the output of the multi-head attention or feed-forward network sub-layer.

Residual connections help address challenges related to training deep networks by allowing the model to learn incremental updates rather than entirely new transformations, which stabilizes training and allows for faster convergence. Additionally, layer normalization further enhances stability by ensuring a consistent distribution of activations across layers.

Input Embeddings

To process textual data, as the original Transformer was designed to do, input tokens are first converted into dense vectors known as embeddings. Each token in the input sequence is mapped to a high-dimensional vector space. This is achieved using an embedding matrix E, where each token t is transformed into its corresponding embedding $E(t)$.

However, the Transformer architecture can also process other types of data sequences. For instance, in the case of multimodal signal data, where each timestep has 6 data signal points, a typical processing step involves employing a learnable linear layer to project the data at each timestep into a higher-dimensional space. This projection allows the model to effectively perform multi-head attention on the input signals.

Positional Encodings

One of the critical challenges in sequence modeling with the Transformer architecture is the lack of inherent mechanisms to capture the order of input tokens. Unlike RNNs, which process tokens sequentially and naturally incorporate positional information through their recurrence mechanisms, Transformers operate on sets of tokens simultaneously. This characteristic makes the Multi-Head Attention block permutation-equivariant, meaning it cannot distinguish whether an input comes before another in the sequence.

In tasks like language understanding, the position of words within a sentence is crucial for accurate interpretation. For example, in the sentence "The cat sat on the mat," the meaning is highly dependent on the order of the words. Therefore, the Transformer model requires a way to integrate positional information into its computations.

To address this, positional encodings are introduced to the input embeddings. These encodings have the same dimension as the embeddings, allowing them to be summed directly with the input features. By adding positional encodings to the input embeddings, the model is provided with information about the position of each token in the sequence, enabling it to consider the order of words during processing.

The positional encoding PE for each position pos and dimension i is defined as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

where d_{model} is the dimension of the embeddings. These sine and cosine functions at different frequencies allow the model to learn the relative positions of the tokens.

2.1.3 Attention Mechanism in the Transformer:

In the Transformer architecture, attention mechanisms are employed in various forms to capture dependencies and relationships within and across sequences. The three main types of attention in the Transformer are self-attention in the encoder, masked self-attention in the decoder, and cross-attention (encoder-decoder attention).

Self-Attention in the Encoder

In the encoder, each layer employs a multi-head self-attention mechanism. Self-attention allows each element in the input sequence to attend to all other elements, enabling the model to capture both local and global dependencies. The process involves computing attention scores for each element with respect to all other elements in the sequence. Since this attention is applied within the same sequence, it is referred to as “self-attention.” The output is a weighted combination of the values based on these attention scores, allowing each element to incorporate information from every other element in the sequence.

In the encoder, there is no restriction on which elements can be attended to; every element can attend to every other element. This enables the model to effectively capture context across the entire input sequence.

Masked Self-Attention in the Decoder

The decoder also uses self-attention, but with an important distinction: it is masked. This masking ensures that when predicting the next element in a sequence, the model only considers the elements processed up to that point and does not see future elements. This is crucial for autoregressive tasks, such as sequence generation, where the model should not have access to future elements that it is supposed to predict.

The mask is implemented by setting the attention scores for future elements to negative infinity, effectively zeroing out the corresponding attention weights after applying the softmax function. This ensures that the model only attends to past elements (and the current one), allowing it to generate sequences in a left-to-right manner.

Cross-Attention

In addition to the masked self-attention mechanism, the decoder has another attention mechanism known as cross-attention. This attention layer allows the decoder to attend to the output of the encoder, integrating information from the input sequence. During the computation of cross-attention, the queries come from the decoder, while the keys and values come from the encoder's output. This mechanism enables the decoder to align and incorporate relevant information from the input sequence while generating the output sequence.

2.1.4 Conclusion

The introduction of the Transformer architecture has revolutionized the way we solve complex problems across many disciplines, particularly within the domain of NLP. The attention mechanism in the Transformer addresses many limitations faced by previous architectures such as RNNs and LSTMs. By eliminating the need for sequential data processing, the Transformer model achieves unprecedented levels of parallelization, significantly improving training efficiency and scalability.

In this chapter, we have explored the fundamental components of the Transformer architecture. The multi-head attention mechanisms and the positional encoding are important components responsible for enhancing the model's ability to capture complex dependencies and for learning contextual information within sequences. Those architectural innovation has set new benchmarks in various NLP tasks and inspired the development of numerous variants and enhancements, such as BERT and GPT.

2.2 Graph Neural Networks

2.2.1 Introduction

In the rapidly evolving field of machine learning, traditional methods like Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance on structured, grid-like data such as images and videos. However, these models struggle with non-Euclidean data types, such as graphs, where the underlying structure is irregular, and relationships between entities are complex and unstructured. This limitation presents significant challenges when applying standard neural network architectures to domains where data is naturally represented as graphs, such as social networks, molecular chemistry, and transportation networks.

Graph Neural Networks (GNNs) have emerged as a powerful solution to these challenges, extending the principles of neural networks to graph-structured data. Unlike CNNs, which rely on a regular grid structure to apply convolutions, GNNs are designed to operate on graphs by leveraging the inherent connectivity and relationships between nodes. This capability allows GNNs to capture complex patterns and dependencies that are crucial in graph-based applications. GNN has proven highly effective in various domains, including molecular property prediction [15, 16, 17], social network analysis [18, 19, 20], computer vision [21, 22], and cybersecurity [23, 24], where understanding the relationships and interactions between entities is paramount.

One of the most prominent types of GNNs is the Graph Convolutional Network (GCN), which generalizes the concept of convolution to graphs. GCNs enable the aggregation of information from a node's neighbors, effectively propagating and transforming features across the graph.

In this section, we begin by introducing the foundamental concepts of graph theory and spectral graph theory, which form the basis for understanding GNNs. Next, we delve into the specifics of GCNs, exploring how they extend traditional neural network methods to effectively process graph-structured data. Finally, we conclude with a summary of the key concepts and insights covered in this section.

2.2.2 Graph Theory

A graph is a fundamental data structure used to model relationships between pairs of objects. Formally, a graph $\mathcal{G}$ is defined as a pair $\mathcal{G} = (V, E)$, where:

- V is a set of nodes (or vertices), defined as $V = \{v_1, v_2, \ldots, v_n\}$.
- E is a set of edges, where $E \subseteq V \times V$. Each edge $e = (v_i, v_j)$ represents a connection or relationship between two distinct nodes v_i and v_j, where $v_i, v_j \in V$.

Depending on how these relationships are represented, graphs can be classified into various types, including:

- **Undirected Graphs:** Edges have no direction, i.e., (v_i, v_j) is identical to (v_j, v_i).
- **Directed Graphs:** Edges have a direction, i.e., (v_i, v_j) is not the same as (v_j, v_i).
- **Weighted Graphs:** Edges have weights, representing the strength or capacity of the connection.
- **Unweighted Graphs:** All edges are treated equally, i.e., the weights are not considered.

2.2.3 Spectral Graph Theory

Spectral Graph Theory provides a way to define convolution on graphs through the eigenvalues and eigenvectors of matrices associated with the graph, such as the adjacency matrix or the Laplacian matrix.

Graph Laplacian

The Graph Laplacian is a matrix representation of a graph that plays a central role in spectral graph theory. For a graph $\mathcal{G}$ with an adjacency matrix $\mathbf{A}$ and a degree matrix $\mathbf{D}$ (where $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$), the unnormalized Laplacian matrix $\mathbf{L}$ is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

The Laplacian matrix $\mathbf{L}$ is symmetric and positive semi-definite, meaning it can be diagonalized as:

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$$

Here:

- $\mathbf{U}$ is an orthogonal matrix containing the eigenvectors of $\mathbf{L}$ as columns.
- $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathbf{L}$.

The eigenvalues and eigenvectors of the Laplacian matrix $\mathbf{L}$ provide a spectrum for the graph, analogous to the Fourier basis in signal processing. The eigenvectors form a basis in which the graph signal (a function defined on the nodes of the graph) can be decomposed, and the eigenvalues provide the frequencies of the corresponding components.

Graph Signal

A graph signal is a function that assigns a value (or set of values) to each node in a graph. Mathematically, if we have a graph $\mathcal{G} = (V, E)$ with n nodes, a graph signal is a vector $\mathbf{x} \in \mathbb{R}^n$ where each element $\mathbf{x}(i)$ corresponds to the signal value at node i.

Graph signals are analogous to time-series signals or image pixels, but instead of being defined over a regular domain (like time or a 2D grid), they are defined over the nodes of a graph. The structure of the graph (i.e., the relationships between nodes encoded in the edges) often provides important contextual information that influences the signal values.

Example of a graph signal: Let's consider a simple social network graph where each node represents a person, and each edge represents a friendship between two people. Suppose we have a scenario where we want to analyze the influence of a certain product's popularity in this social network. A graph signal in this context could represent the level of interest or opinion about the product, with each node's signal value indicating how much interest the corresponding person has.

Let's construct a simple example with a graph $\mathcal{G}$ of 4 nodes:

- **Graph Structure:**
  - Nodes: $V = \{1, 2, 3, 4\}$
  - Edges: $E = \{(1, 2), (2, 3), (3, 4)\}$
  - Adjacency Matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

- **Graph Signal Representation:**
  - Suppose the signal $\mathbf{x}$ represents how interested each person is in a new product, with values ranging from 0 (no interest) to 10 (very interested).
  - Let's say the signal vector $\mathbf{x}$ is:

$$\mathbf{x} = \begin{pmatrix} 7 \\ 5 \\ 2 \\ 4 \end{pmatrix}$$

  - Here:
    - Node 1 has a signal value of 7 (high interest).
    - Node 2 has a signal value of 5.
    - Node 3 has a signal value of 2 (low interest).
    - Node 4 has a signal value of 4.

This graph signal $\mathbf{x}$ represents the interest levels of the individuals in the social network. The graph structure (encoded in the adjacency matrix $\mathbf{A}$) can help us understand how interest might spread or influence other nodes in the network, which is where graph signal processing and GCNs come into play.

When we have more than one feature associated with each node in the graph, we have a multidimensional graph signal. For instance, if each node in our social network graph also has an interest level for a second product, the graph signal would be a matrix where each row represents a node and each column represents a feature (e.g., interest in Product A and Product B).

Graph Fourier Transform

The Graph Fourier Transform allows us to analyze a graph signal in the spectral domain, analogous to the Fourier Transform in signal processing. The Graph Fourier Transform of a graph signal $\mathbf{x} \in \mathbb{R}^n$ is defined as:

$$\hat{\mathbf{x}} = \mathbf{U}^\top \mathbf{x}$$

Where $\mathbf{U}$ is the matrix of eigenvectors of the Laplacian matrix $\mathbf{L}$. This operation transforms the signal $\mathbf{x}$ into the spectral domain, where $\hat{\mathbf{x}}$ represents the coefficients of the signal in the basis of the eigenvectors of the Laplacian.

The inverse Graph Fourier Transform reconstructs the signal from its spectral components:

$$\mathbf{x} = \mathbf{U}\hat{\mathbf{x}}$$

Spectral Graph Convolution

The convolution of a graph signal $\mathbf{x}$ with a filter $g(\mathbf{L})$ in the spectral domain is defined as:

$$\mathbf{y} = g(\mathbf{L})\mathbf{x} = \mathbf{U}g(\mathbf{\Lambda})\mathbf{U}^\top \mathbf{x}$$

Here:

- $g(\mathbf{L})$ is a function of the Laplacian matrix, which acts as a filter.
- $g(\mathbf{\Lambda})$ is a diagonal matrix where $g(\mathbf{\Lambda})_{ii} = g(\lambda_i)$, i.e., the filter function applied to each eigenvalue.

This operation is analogous to applying a filter in the Fourier domain in classical signal processing.

Chebyshev Polynomial Approximation

To avoid the computational burden of explicitly computing the eigenvectors $\mathbf{U}$ and the filter function $g(\mathbf{\Lambda})$, Kipf et al. [3] proposed to approximated the spectral filter $g(\mathbf{L})$ using Chebyshev polynomials. Chebyshev polynomials can be defined as follows:

$$g(\mathbf{L}) \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\mathbf{L}})$$

Where:

- θ_k are coefficients to be learned during training.
- $T_k(\tilde{\mathbf{L}})$ are the Chebyshev polynomials evaluated at the scaled Laplacian $\tilde{\mathbf{L}}$.

Chebyshev polynomials are defined recursively as:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \quad \text{for } k \geq 2$$

In the context of graph convolution, x is replaced by the scaled Laplacian $\tilde{\mathbf{L}}$:

- For $k = 0$: $T_0(\tilde{\mathbf{L}}) = \mathbf{I}$ (the identity matrix).
- For $k = 1$: $T_1(\tilde{\mathbf{L}}) = \tilde{\mathbf{L}}$.
- For $k = 2$ and higher: $T_k(\tilde{\mathbf{L}})$ involves higher-order polynomials of $\tilde{\mathbf{L}}$.

Transition to the Simplified GCN

1. **First-Order Approximation ($K = 1$):**
   - The GCN simplifies the Chebyshev approximation by considering only the first two terms in the series (i.e., a first-order approximation):
   $$g(\mathbf{L}) \approx \theta_0 T_0(\tilde{\mathbf{L}}) + \theta_1 T_1(\tilde{\mathbf{L}})$$
   - Substituting $T_0(\tilde{\mathbf{L}}) = \mathbf{I}$ and $T_1(\tilde{\mathbf{L}}) = \tilde{\mathbf{L}}$, we get:
   $$g(\mathbf{L}) \approx \theta_0 \mathbf{I} + \theta_1 \tilde{\mathbf{L}}$$
2. **Renormalization Trick:**
   - Instead of directly using the scaled Laplacian $\tilde{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, the GCN formulation simplifies it by introducing a "renormalization trick." The idea is to approximate the operation of $\tilde{\mathbf{L}}$ by working directly with the adjacency matrix. The adjacency matrix is adjusted to include self-loops:

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$$

The corresponding degree matrix is:

$$\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$$

The normalized adjacency matrix is then defined as:

$$\tilde{\mathbf{A}}_{\text{norm}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$$

- This normalized adjacency matrix $\tilde{\mathbf{A}}_{\text{norm}}$ acts as a simplified substitute for the scaled Laplacian $\tilde{\mathbf{L}}$.

3. **Final GCN Expression:**

- The final GCN layer can then be expressed as:

$$\begin{aligned}\mathbf{H}^{(l+1)} &= \sigma\left(\tilde{\mathbf{A}}_{\text{norm}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \\ &= \sigma\left(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)\end{aligned}$$

- where $\tilde{\mathbf{A}}_{\text{norm}}$ effectively plays the role of the simplified filter $g(\mathbf{L})$ from the spectral domain, but computed in the spatial domain using normalized adjacency. $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with added self-loops (identity matrix $\mathbf{I}$), $\tilde{\mathbf{D}}$ is the degree matrix corresponding to $\tilde{\mathbf{A}}$, with $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$, $\mathbf{W}^{(l)}$ is the learnable weight matrix for layer l, and σ is a non-linear activation function, such as ReLU.

4. **Importance of Normalization:**

- Normalization of the adjacency matrix is crucial for the stability of the graph convolution operation. Without normalization, the features of nodes with a high degree would dominate the aggregation process, leading to numerical instability and poor model performance. The normalization ensures that the convolution operation remains stable and that all nodes contribute equally to the aggregation process, regardless of their degree.
- This normalization is motivated by the desire to make the convolution operation invariant to the degree of the nodes. Consider the propagation of features for a single layer:

$$\mathbf{H}^{(l+1)} = \tilde{\mathbf{A}}_{\text{norm}}\mathbf{H}^{(l)}\mathbf{W}^{(l)} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l)}\mathbf{W}^{(l)}$$

- Each element $(\mathbf{H}^{(l+1)})_i$ of the feature matrix $\mathbf{H}^{(l+1)}$ is computed as:

$$(\mathbf{H}^{(l+1)})_i = \sum_j \frac{1}{\sqrt{d_i d_j}}\tilde{A}_{ij}(\mathbf{H}^{(l)})_j\mathbf{W}^{(l)}$$

- Where d_i and d_j are the degrees of nodes i and j in the graph. This ensures that the contribution of each neighboring node j to node i's new features is scaled by the square root of the product of their degrees, leading to a balanced aggregation of features.

2.2.4 Spatial-based GCNs

Spatial-based GCNs define graph convolutions directly in the node's neighborhood. The operation involves aggregating feature information from neighboring nodes (and possibly the node itself) and then applying a transformation.

The general update rule for a spatial-based GCN is:

$$H_i^{(l+1)} = \sigma\left(W^{(l)} \cdot \text{AGGREGATE}\left(\left\{\frac{1}{c_{ij}} H_j^{(l)} : j \in \mathcal{N}(i) \cup \{i\}\right\}\right)\right)$$

where:

- $H_i^{(l)}$ is the feature vector of node i at layer l. This vector represents the current state or features of node i before the update at layer $l+1$.
- $\mathcal{N}(i)$ denotes the set of neighbors of node i. This is the set of nodes that are directly connected to node i by an edge.
- c_{ij} is a normalization constant, often chosen as $\sqrt{d_i d_j}$ where d_i and d_j are the degrees of nodes i and j, respectively. This normalization helps to balance the influence of nodes based on their degree, preventing features from being disproportionately influenced by nodes with many neighbors.
- $W^{(l)}$ is the weight matrix for the l-th layer. This matrix contains the learnable parameters that will transform the features from the previous layer to the current layer.
- AGGREGATE is the aggregation function. This function takes the set of transformed feature vectors from the neighbors (and possibly the node itself) and combines them into a single vector. Common choices for aggregation include:
  - **Summation**: Simply adds the feature vectors together.
  - **Mean**: Takes the average of the feature vectors.
  - **Max Pooling**: Selects the maximum value for each feature across all vectors.
  - **Attention-based Aggregation**: Uses learned attention weights to combine the feature vectors.
- σ is a non-linear activation function, such as ReLU (Rectified Linear Unit) or sigmoid. This function introduces non-linearity into the model, allowing it to capture more complex patterns in the data.

In the case where the AGGREGATE function in the spatial-based method is a weighted sum (with weights derived from normalized adjacency matrix entries), the resulting spatial-based GCN can be mathematically equivalent to the spectral-based GCN derived from the spectral approximation of [3].

Specifically, if we use the normalized adjacency matrix $\tilde{A} = D^{-1/2}(A + I)D^{-1/2}$ in the aggregation, the spatial-based GCN's propagation rule:

$$H^{(l+1)} = \sigma\left(\tilde{A}H^{(l)}W^{(l)}\right)$$

is identical to the propagation rule derived from the spectral approximation.

2.2.5 Applications of GNNs

GNNs have found a lot of applications across a wide range of tasks, including but not limited to the following:

Node Classification: Node classification involves predicting the label of a node based on both its features and the structure of the graph. GNNs are particularly effective for this task as they aggregate information from neighboring nodes, enabling a more holistic understanding of each node's role within its graph.

Graph Classification: Graph classification involves predicting the label of an entire graph based on its structure and the features of its nodes. GNNs are well-suited for this task as they can capture both global and local graph patterns.

Link Prediction: Link prediction aims to predict the existence or strength of a connection between two nodes. Examples include recommending friends in social networks, identifying missing or potential interactions in biological networks, and suggesting relevant items in recommendation systems.

Graph Clustering: Graph clustering involves partitioning a graph into groups (clusters) of nodes that share similar properties or roles. This task is critical in community detection within social networks, grouping proteins with similar functions in biological networks, or clustering users in recommendation systems.

2.2.6 Conclusion

Graph Neural Networks represent a significant advancement in deep learning, enabling the processing of graph-structured data in a manner similar to CNNs on images. By leveraging the spectral properties of graphs or directly operating in the spatial domain, GNNs can capture complex relationships and patterns in non-Euclidean domains.

The development of GNNs is rooted in spectral graph theory, where the graph convolution operation is derived from the eigenvalues and eigenvectors of the Laplacian matrix. This allows GNNs to generalize the concept of convolution to graphs, enabling their application to a wide range of tasks.

2.3 Multimodal Machine Learning

2.3.1 Introduction

Multimodal Machine Learning (MML) is an emerging area in AI that focuses on the integration and processing of data from multiple modalities. A modality refers to a particular type of data representation, such as text, images, audio, video, or sensor data. MML aims to leverage complementary information across different modalities to achieve more accurate, robust, and generalizable models than those using a single modality.

Inspired by the human brain's ability to integrate multiple sensory inputs, MML seeks to replicate this processing in machines. The human brain is inherently multimodal, processing and integrating information from various sensory inputs—such as sight, sound, touch, taste, and smell—simultaneously. This ability allows humans to develop a rich, coherent understanding of the world. For example, when listening to someone speak, the brain processes the auditory information (the words being spoken) along with visual information (the speaker's facial expressions and gestures) to interpret the full meaning of the communication. MML aims to replicate this human-like ability in machines, enabling them to leverage diverse data sources for more accurate and robust decision-making.

The potential of MML has already been demonstrated across a variety of tasks, ranging from emotion analysis [25, 26] and visual question answering [27, 28] to surgical gesture recognition [29, 30, 31]. These achievements highlight how MML's integration of different data modalities can push the boundaries of unimodal model.

2.3.2 Motivation

While the successes of unimodal learning approaches have driven significant advancements in AI, real-world data is often multimodal by nature. Single-modality models are limited in their ability to fully capture the complexity and richness of this data, which can result in models that are less accurate, less generalizable, and prone to errors when confronted with diverse, dynamic environments. For instance, in human-computer interaction, relying solely on text input ignores crucial information from visual cues like facial expressions and gestures. In autonomous driving, processing only camera images can miss critical data from other modalities like LIDAR, radar, and vehicle dynamics, leading to safety risks. MML directly addresses these limitations by integrating complementary information from multiple sources, thereby enhancing the contextual understanding and decision-making capabilities of models.

2.3.3 Multimodal Learning Tasks

MML involves several fundamental research themes that address how different modalities are represented, aligned, and integrated to build effective multimodal models. These research themes are critical for understanding the complexities of learning from multi-

ple information sources and designing systems that can handle the inherent challenges of multimodal data. In this section, we explore the main research themes: Representation Learning, Alignment, Co-Learning, and Multimodal Fusion.

Representation Learning

Representation learning in MML focuses on how to encode information from multiple modalities into meaningful and unified representations. The challenge lies in capturing relevant information from each modality while accounting for their diverse structures, dynamics, and scales. In multimodal systems, representation learning typically revolves around two primary approaches: shared representations and modality-specific representations.

In shared representation learning, the goal is to map different modalities into a common feature space where their features can be directly compared. By learning a joint embedding space, the model can capture the correlations and complementarities between the modalities, leading to richer and more robust representations. Canonical Correlation Analysis (CCA) [32] and its deep learning extension (Deep CCA) [33] are commonly used to learn such shared spaces by maximizing the correlation between modalities.

On the other hand, modality-specific representations maintain separate feature spaces for each modality, allowing for more specialized processing that preserves the unique characteristics of each data source. In such systems, cross-modal information sharing can be achieved through mechanisms like attention layers in cross-modal transformers [34], enabling the integration of information while retaining modality-specific details. This method is advantageous in scenarios where the modalities are structurally different or where retaining distinct information from each modality is crucial for the task at hand.

Alignment

Alignment is a crucial task in MML, focusing on the need to establish correspondences between different modalities that may have distinct temporal, spatial, or structural characteristics. Misalignment can occur due to differences in timing, sampling rates, or data structures, making it essential to address these discrepancies for effective integration.

Temporal alignment is particularly important in tasks involving time-dependent data, such as video and audio. For instance, synchronizing visual frames with corresponding audio signals is a common requirement in applications like video analysis and speech recognition. Ensuring that these sequences are properly aligned allows the model to integrate information across modalities in a coherent and consistent manner. Techniques such as attention mechanisms [34] and contrastive learning [35] are often employed to address these challenges.

Co-Learning

Co-learning addresses how knowledge and signals from one modality can improve learning in another, particularly in scenarios where one modality might have limited data or noisy information. Co-learning is essential for enhancing the robustness and generalization capabilities of multimodal models, especially in real-world environments where data quality and availability can vary significantly across modalities.

Common co-learning strategies include co-training [36], where models iteratively refine each other by exchanging predictions; transfer learning [37], where knowledge is transferred across modalities; and multi-task learning [38], where a single model is trained on multiple tasks involving different modalities. These strategies aim to make models more resilient and generalizable, especially in real-world applications where data may be incomplete or noisy.

Multimodal Fusion

Multimodal fusion is the process of integrating information from multiple modalities to create more comprehensive models. Fusion can occur at various stages of a model's architecture and can be broadly categorized into three strategies: early fusion, late fusion, and intermediate fusion. Early fusion combines features at the input level, allowing the model to learn joint representations from the outset. Late fusion processes each modality independently and combines their outputs at the decision level, which is useful when modalities have distinct characteristics. Intermediate fusion fused modality-specific features at a later stage, balancing the flexibility of late fusion with the comprehensiveness of early fusion.

In addition to these strategies, there exist more complex fusion techniques that can be applied at any of these stages, such as tensor fusion [39, 40], attention-based fusion [41, 42], and graph-based fusion [43], which aim to capture higher-order interactions and more nuanced relationships across modalities. These advanced techniques push beyond simple concatenation or averaging, enabling richer and more expressive multimodal representations. These fusion techniques will presented in more details in the following.

2.3.4 Multimodal Fusion

Fusion of multimodal data is a central problem in MML, and several strategies have been developed to address it. These strategies can be broadly categorized into early fusion, late fusion, intermediate fusion, tensor fusion, attention mechanisms, and graph-based fusion.

Early Fusion

Early fusion, also known as feature-level fusion, involves merging features extracted from different modalities at an early stage, typically before feeding them into a learning algorithm. Mathematically, suppose we have two modalities, $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$, where d_1

and d_2 are the dimensions of the feature spaces. In early fusion, the features from the two modalities are concatenated to form a combined feature vector:

$$x_{\text{fused}} = [x_1; x_2] \in \mathbb{R}^{(d_1+d_2)}$$

This fused feature vector x_{fused} is then used as input to a learning algorithm, such as a neural network or a classifier. Early fusion allows the model to learn joint representations of the data, but it can suffer from high dimensionality and difficulties in capturing complex interactions between modalities, especially when the feature spaces of the modalities are very different.

Late Fusion

Late fusion, also known as decision-level fusion, involves processing each modality separately and then combining the decisions or outputs from each modality. Let $f_1(x_1)$ and $f_2(x_2)$ be the outputs from models trained on each modality separately. The final decision y can be obtained by combining these outputs, often using a weighted sum or a voting mechanism:

$$y = g(f_1(x_1), f_2(x_2))$$

Common choices for $g(\cdot)$ include averaging, weighted sum, voting mechanisms, or more complex ensemble methods. Late fusion is simpler and can effectively handle heterogeneous modalities, but it might miss potential interactions between modalities that could be captured in earlier stages of processing.

Intermediate Fusion

Intermediate fusion combines elements of both early and late fusion. In hybrid fusion, features from different modalities are partially fused at intermediate layers before final decision-making.

Let $h_1(x_1) \in \mathbb{R}^{m_1}$ and $h_2(x_2) \in \mathbb{R}^{m_2}$ represent the intermediate feature embeddings learned from each modality. Intermediate fusion involves combining these embeddings:

$$x_{\text{fused}} = g(h_1(x_1), h_2(x_2))$$

where $g(\cdot)$ can be any function that fuses the learned embeddings, such as concatenation or element-wise addition. This fused representation x_{fused} is then passed through further layers before making a final prediction. Intermediate fusion aims to balance the benefits of both early and late fusion, capturing inter-modal interactions while maintaining modularity in the model design.

Tensor Fusion

Tensor fusion expands feature vectors from different modalities into multi-dimensional tensors to capture higher-order interactions between the modalities. Unlike simpler approaches that only consider linear combinations of features (like concatenation or element-wise addition), tensor fusion allows for the modeling of multiplicative interactions between features across modalities. Consider two modalities with feature vectors $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$. In tensor fusion, the outer product $\otimes$ is used to combine these feature vectors into a tensor:

$$\mathbf{T} = x_1 \otimes x_2 \in \mathbb{R}^{d_1 \times d_2}$$

The resulting tensor $\mathbf{T}$ represents all pairwise interactions between elements of x_1 and x_2, forming a 2D matrix where each entry is calculated as:

$$\mathbf{T}_{ij} = (x_1)_i \cdot (x_2)_j$$

This can be generalized for more than two modalities. For three modalities with features $x_1 \in \mathbb{R}^{d_1}$, $x_2 \in \mathbb{R}^{d_2}$, and $x_3 \in \mathbb{R}^{d_3}$, the resulting tensor $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ captures all triplet interactions:

$$\mathbf{T}_{ijk} = (x_1)_i \cdot (x_2)_j \cdot (x_3)_k$$

Tensor fusion goes beyond linear interactions, capturing complex, multiplicative relationships between features from different modalities. The resulting tensor contains richer and more expressive representations that incorporate higher-order correlations across modalities, which can lead to more accurate and nuanced decision-making. The primary drawback of tensor fusion is the exponential growth in dimensionality. For instance, for two feature vectors with dimensions d_1 and d_2, the resulting tensor will have $d_1 \times d_2$ elements. This can quickly become computationally prohibitive, especially when dealing with more than two modalities.

Attention Mechanisms

Attention-based models dynamically weigh the contributions of different modalities depending on the context or task at hand. Mathematically, attention mechanisms can be represented as a weighted sum of modality-specific features. For a given modality $\mathbf{X}_m$, an attention weight α_m is computed based on the relevance of the modality to the task:

$$\alpha_m = \frac{\exp(e_m)}{\sum_k \exp(e_k)}$$

where e_m is an alignment score computed using a function that measures the compatibility between the modality and the task, often using a neural network. The final multimodal representation $\mathbf{X}_{\text{att}}$ is a weighted sum of the modality-specific features:

$$\mathbf{X}_{\text{att}} = \sum_m \alpha_m \mathbf{X}_m$$

Attention mechanisms allow the model to focus on the most relevant parts of each modality, leading to more flexible and accurate multimodal representations.

Graph-based Fusion

In cases where relationships between modalities are complex and structured, graph-based fusion techniques can model these relationships explicitly. A graph $\mathcal{G} = (V, E)$ is constructed where nodes V represent different modalities or features, and edges E capture the dependencies between them. The feature representation of each node can be updated using a message-passing mechanism, where information is aggregated from neighboring nodes:

$$\mathbf{H}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \mathbf{W}^{(l)} \mathbf{H}_j^{(l)}\right)$$

where $\mathbf{H}_i^{(l)}$ is the embedding of node i at layer l, $\mathcal{N}(i)$ represents the neighbors of node i, and $\mathbf{W}^{(l)}$ is the weight matrix at layer l. This approach is particularly useful when the relationships between modalities are complex and need to be explicitly modeled. Furthermore, GNNs can efficiently handle a large number of modalities and their interactions. However, some limitations need to be aknowledge. Defining the graph structure (e.g., which nodes are connected) can be non-trivial and often requires domain-specific knowledge. Additionally, Deep GNNs can suffer from over-smoothing, where all node embeddings become indistinguishable after many layers.

2.3.5 Challenges

Despite its promise, MML faces several challenges that complicate the integration and processing of multimodal data:

Heterogeneity of Data: Different modalities often have distinct data structures, formats, and statistical properties. For instance, text data is sequential and discrete, while image data is spatial and continuous. This heterogeneity makes it challenging to develop unified models that can effectively process and integrate diverse types of data.

Alignment of Modalities: Aligning data from different modalities is a significant challenge. For example, in video analysis, aligning spoken words with corresponding visual cues requires precise temporal synchronization. Misalignment can lead to incorrect or incomplete interpretation of the data.

Fusion Strategies: Determining the optimal strategy for fusing information from different modalities is non-trivial. Simple concatenation of features might not capture complex interactions between modalities, while more sophisticated fusion techniques might require significant computational resources and may introduce noise or redundancy.

Missing Data: In many practical scenarios, not all modalities are available at all times. Handling missing data and ensuring the model remains robust in such situations is a key challenge in MML.

Scalability: MML systems often require large amounts of data and computational power, particularly when dealing with high-dimensional data such as images and videos. Ensuring that MML models scale efficiently with increasing data complexity and volume is crucial.

Interpretability: As with many machine learning models, interpretability remains a challenge. Understanding how different modalities contribute to the final decision-making process is important, particularly in applications where transparency is critical, such as healthcare.

2.3.6 Conclusion

Multimodal Machine Learning represents a significant step forward in the development of intelligent systems capable of processing and understanding complex, heterogeneous data. By integrating information from multiple modalities, MML models can achieve superior performance across a wide range of tasks. However, the field is still in its early stages, and numerous challenges remain, particularly in areas such as data fusion, representation learning, and model interpretability.

Part I

Multimodal Affective Computing

Chapter 3

Affective Computing

Contents

This chapter provides the theoretical background on the field of affective computing, offering the necessary context for the methods introduced in Chapters 4 and 5.

Section 3.1 introduces the field of affective computing and its associated challenges. Next, Section 3.2 discusses the role of emotion in daily life, the categorical and dimensional emotion models, and the various modalities through which emotions are expressed. Finally, Section 3.3 provides an overview of emotion recognition methods, covering both unimodal and multimodal approaches employing these modalities.

3.1 Introduction

Affective Computing is a multidisciplinary field at the intersection of computer science, psychology, and cognitive science, aiming to develop systems and devices that can recognize, interpret, express, and process human emotions. The importance of Affective Computing lies in its potential to revolutionize how we interact with technology. By incorporating emotional intelligence into devices and software, technology can become more responsive to the nuances of human moods, stress levels, and emotional needs. This can enhance user experience in a wide range of applications, from personalized learning and mental health monitoring to customer service and entertainment.

For instance, educational software that adapts to a student's frustration or boredom can offer alternative explanations, motivating them to persevere. In mental health, wearables that detect stress or anxiety levels can prompt users to take a break or engage in a calming activity. Moreover, in customer service, chatbots and virtual assistants that understand and respond to a user's emotional state can provide more empathetic and effective support.

Affective Computing leverages machine learning and pattern recognition techniques for processing emotional-based signals , in order to recognize and classifying emotion states. These emotion-based signals can include facial expressions, body movements, physiological signals, speech, or text data. However, emotion recognition is inherently challenging due to the subjective and context-dependent nature of emotions. Variations in emotional expression across cultures, individuals, and situations make accurate interpretation difficult. For instance, emotions like anxiety or frustration may be concealed or exaggerated, complicating detection.

Moreover, human emotions are also complex, subtle, and multifaceted, often overlapping and difficult to categorize. For example, a person may exhibit signs of multiple emotions simultaneously, further complicating the process of emotional labeling.

In recent years, multimodal approaches have become increasingly popular. These methods combine diverse emotional cues, such as facial expressions, speech patterns, and physiological data (e.g., heart rate or skin conductance), to improve emotion recognition accuracy. The integration of physiological signals is especially valuable as they provide objective, involuntary indicators of emotional states, offering insights that external expressions might obscure.

Nevertheless, integrating multiple data modalities presents its own set of challenges.

Differences in temporal resolution and data structures can complicate the alignment of data across modalities. Precise synchronization and integration require efficient algorithms and models capable of real-time emotion recognition. One significant challenge is the effective fusion of various modalities. Developing advanced fusion techniques tailored to the specific modalities involved is crucial for maximizing the potential of each modality and improving overall recognition performance.

3.2 Theoretical Background

3.2.1 Human Emotions

Emotions play a crucial role in our human experience, impacting nearly every aspect of our lives. They influence our decisions and actions, with positive emotions guiding us towards beneficial activities, relationships, or goals, while negative emotions often serve as signals to avoid challenging situations. Emotions are also central to human communication and can be expressed through facial expressions and body language, allowing us to convey feelings nonverbally.

Moreover, emotional well-being is closely linked to physical health. Chronic stress, anxiety, and depression can manifest physically, increasing the risk of conditions such as heart disease, diabetes, and weakened immune function. Conversely, positive emotions and emotional resilience can improve physical health and overall well-being.

The complexity and variability of emotional experiences highlight their deeply personal nature, influenced by a dynamic interplay of cultural, genetic, and experiential factors. Cultural norms, for instance, shape how emotions are expressed and perceived—some cultures emphasize emotional restraint, while others encourage openness and expressiveness [44]. Genetic predispositions can affect emotional reactivity and vulnerability to mood disorders, revealing the biological foundations of emotions [45]. Additionally, personal experiences, including trauma and life achievements, further shape our emotional responses and landscapes. This diversity in emotional experiences underscores that, while emotions are universal, their expression and impact are highly individualized, reflecting a wide range of human diversity.

3.2.2 Emotion models

The study of emotions has led to the development of various models to categorize and understand emotional experiences. These models are broadly classified into two categories: categorical models and dimensional models. Both have contributed significantly to the field of emotion research, including the development of emotion recognition technology.

Categorical Models

Categorical models of emotions propose that a limited number of distinct and universal emotions exist. These models are grounded in the theory that certain emotions are biologically and psychologically fundamental to all humans, regardless of cultural differences [46, 47]. The most influential work in this area is by Paul Ekman [47], a pioneer in the study of emotions and their relation to facial expressions. Paul Ekman identified six basic emotions that he argued were universally recognized and expressed by specific facial expressions: happiness, sadness, fear, disgust, anger, and surprise. Ekman's cross-cultural studies demonstrated that people from diverse cultures could accurately identify these basic emotions from facial expressions. His work led to the development of the Facial Action Coding System (FACS) [48], a comprehensive tool for categorizing the physical expression of emotions through facial movements.

More recently, Cordaro and Keltner, former students of Ekman, have conducted research suggesting an expansion of the list of universal emotions [49]. Their cross-cultural study provides evidence for the universal recognition of additional emotions, including amusement, awe, contentment, desire, embarrassment, pain, and relief, through both facial and vocal expressions.

Despite these advancements, categorical models have limitations. By focusing on a finite set of basic emotions, these models may oversimplify the complexity of human emotional experiences, potentially overlooking the broader spectrum of emotions that individuals experience.

Dimensional Models

Dimensional models, on the other hand, view emotions as existing along continuous dimensions rather than as discrete categories. Russell proposed the circumplex model of emotion [50], which represents emotions on a two-dimensional circular space. Valence is represented on the horizontal axis, with emotions ranging from displeasure to pleasure; it measures how positive or negative an emotion is. Arousal is represented on the vertical axis, evaluating the level of arousal associated with an emotion, measuring the intensity of an emotion, from low to high. Emotions opposite each other on the circle have opposite valence and/or arousal characteristics. For example, happiness (high valence, high arousal) is opposite to sadness (low valence, low arousal). Furthermore, emotions close to each other on the circle are similar in nature. For instance, both joy and surprise might share a high arousal characteristic but differ in their valence.

3.2.3 Emotions related Modalities

Behavior Modalities

Behavioral cues play an important role in how emotions are externally manifested and interpreted. Humans express emotions through several sensory channels, including facial

expressions and body movements. These non-verbal and paralinguistic signals are key for emotional communication and provide crucial data for emotion recognition systems.

Facial Action Units (FAUs): One of the most comprehensive taxonomy for understanding facial expressions is the Facial Action Coding System (FACS), introduced by Ekman and Friesen [48]. FACS defined a comprehensive set of individual muscles or groups of muscles, known as action units (AUs). By combining AUs, it is possible to encode any facial expression, enabling the inference of an individual's emotional state. For instance, AU12 is related to a lip corner pull, typically associated with smiling, while AU4 (brow lowerer) is linked to sadness or anger.

Tracking AUs involves both manual and automated techniques. Manual detection requires trained coders to analyze facial movements from images or videos, which can be labor-intensive and susceptible to inter-coder variability. This method, while accurate, often struggles with subtle and fleeting expressions and requires significant time investment. In contrast, automated techniques utilize machine learning and computer vision [51, 52, 53] algorithms to detect AUs in real-time, enhancing efficiency and scalability. These systems, however, are not without their challenges; they may be affected by variations in lighting, occlusions, or differences in individual facial anatomy, which can impact the accuracy of AU detection.

Facial Landmarks: Facial landmarks are key points on the face (e.g., the corners of the eyes, mouth, and nose) that capture the geometry of facial expressions. These points can be tracked and measured over time to assess the intensity and dynamics of facial expressions.

The tracking of facial landmarks has been made possible by advancements in computer vision. Automated methods, including sophisticated algorithms and deep learning models [54, 55, 56], enable real-time detection and tracking of these landmarks with high precision. However, these methods can encounter difficulties in the presence of lighting variations, facial occlusions, and diverse head poses, which may affect the robustness and reliability of landmark detection.

Body Movements and Posture: In addition to facial expressions, emotions can also be expressed through body movements and posture. For instance, slumped shoulders may indicate sadness, while an upright posture can convey confidence or happiness. Emotion recognition systems can leverage body movement tracking to detect emotions, with the advantage that body cues often reflect emotions even when facial expressions are suppressed.

Tracking body movements has traditionally involved motion capture systems, which provide precise and detailed data but can be cumbersome and impractical for everyday applications. Recent developments in computer vision, such as pose estimation models [57], offer a more practical solution by detecting and analyzing body skeletons in real-time. These methods have made significant strides in accuracy and ease of use, yet they still face challenges such as sensitivity to environmental conditions and the need for substantial computational resources.

Behavioral signals are critical for emotion recognition systems because they reflect how emotions are consciously or unconsciously expressed in social contexts. The challenge, however, is that behavioral cues can be intentionally masked or altered, which can reduce the accuracy of emotion recognition systems that rely solely on these cues.

In the context of one of our primary objectives, which is monitoring the stress levels and emotional states of medical students during medical simulation training, the real-time analysis of facial action units, facial landmarks, and body movements will enable precise emotion recognition. For body movement analysis, using a 3-axis accelerometer (ACC) will provide a practical and unobtrusive method for capturing dynamic body movements in real-world scenarios. The accelerometer measures acceleration along three axes (x, y, and z), enabling detailed tracking of various body movements and postural changes.

Physiological Signals

In addition to behavioral manifestations, emotions also produce distinct physiological changes in the body. These signals, often less conscious and more difficult to control, offer an objective measure for emotion recognition. The physiological changes associated with emotions are grounded in the autonomic nervous system (ANS), the endocrine system, and the central nervous system (CNS).

Emotions trigger specific physiological responses as part of the body's fight-or-flight mechanism. These changes are controlled by the sympathetic and parasympathetic branches of the ANS. The Sympathetic Nervous System (SNS) is responsible for preparing the body to respond to stressful or dangerous situations, often referred to as the "fight-or-flight" response. When activated, the SNS increases heart rate, dilates pupils, enhances blood flow to muscles, and releases stress hormones like adrenaline, all of which prime the body for action [58].

In contrast, the Parasympathetic Nervous System (PNS) works to calm the body and promote relaxation once the threat has passed. Often described as the "rest-and-digest" system, the PNS decreases heart rate, conserves energy, and facilitates digestion and recovery, helping the body return to a balanced state [58].

The following physiological signals are commonly involved in emotion expression:

- **Electrodermal Activity (EDA):** Measures the electrical conductance of the skin, which increases with sweat gland activity during emotional arousal [59].
- **Electrocardiogram (ECG):** Measures the electrical activity of the heart over time. It provides detailed information on heart rate and the patterns of heartbeats. This measurement is a key indicator of emotional expression.
- **Blood Volume Pulse (BVP):** Measures the blood flow through the peripheral blood vessels using a photoplethysmograph (PPG) sensor, often placed on the finger or wrist. Changes in BVP indicate variations in vascular constriction and dilation, which are influenced by emotional states.

- **Respiration (RESP):** This is the number of breaths taken per minute. Increases in respiratory rate are commonly associated with arousal states such as anxiety, fear, or excitement. Conversely, a decreased respiratory rate can indicate relaxation or a calm state. Respiratory rate can be measured using sensors such as respiratory belts or wearable devices equipped with respiratory sensors.

- **Skin Temperature (TEMP):** Variations in skin temperature can reflect changes in the autonomic nervous system's activity, which is closely linked to emotional responses. This is because the ANS affects blood flow and sweat production, both of which influence skin temperature. Several methods can be employed to measure skin temperature, including thermal imaging and contact-based sensor.

- **Electroencephalography (EEG):** Measures electrical activity in the brain, providing insights into a person's emotional state by analyzing brainwave patterns.

Regarding the objective of this thesis to develop stress detection and emotion recognition models for monitoring medical students' stress levels and emotional states during medical simulation scenarios, all the physiological signals mentioned, except EEG, can be monitored for this purpose. EEG signals require invasive equipment, which limits their practicality and usability in real-time, non-intrusive applications. In contrast, the other physiological signals can be monitored using less invasive wearable sensors, enhancing their feasibility and applicability in such settings.

3.3 Emotion Recognition

Emotion recognition is the process of identifying and classifying human emotions based on various data modalities, including behavioral cues (e.g., facial expressions, body movements) and physiological signals (e.g., heart rate, skin conductance). This section provides an overview of emotion recognition methods based on unimodal and multimodal approaches.

3.3.1 Behavior Modalities

Facial expressions are one of the most commonly used modalities for emotion recognition. Traditional techniques often relied on manually crafted features and conventional machine learning methods. For instance, Pu et al. [60] introduced a two-stage facial expression recognition framework. The first stage involved detecting Facial Action Units (AUs) using a Random Forest classifier trained on features extracted from the Active Appearance Model (AAM) and optical flow computations. The second stage involved feeding the detected AUs as inputs to another Random Forest classifier for facial expression recognition, achieving a 96.38% recognition rate on the Cohn–Kanade (CK+) dataset [61]. Kumar et al. [62] developed a system that initially detects faces and identifies facial landmarks. Using these landmarks, the system extracts facial patches and computes Histogram of Oriented

Gradients (HOG) features for each patch. These features are then concatenated and classified using a Support Vector Machine (SVM) for facial expression recognition on the CK+ dataset. Abdulrahman et al. [63] proposed a method that extracts features using the Local Binary Pattern (LBP) technique and applies Principal Component Analysis (PCA) for dimensionality reduction, followed by a SVM classifier.

Despite their success, these traditional methods rely heavily on handcrafted features and can be constrained by the complexity of feature extraction and the need for domain expertise. Recent advances in deep learning have shifted the focus towards automatic feature learning. For example, Ouellet [64] combined a pre-trained CNN from ImageNet [65] with an SVM classifier for facial expression recognition, achieving 94.40% accuracy across seven emotion classes on the CK+ dataset. Similarly, Li et al. [66] employed a ResNet-50 [67] architecture for facial expression recognition. In [68], the authors proposed a two-stream framework for video-based emotion recognition. Their approach integrated a Recurrent Neural Network (RNN) with sequences of image features extracted from the VGG16 model [69], alongside a 3D Convolutional Neural Network (CNN), and an audio module that extracted audio features and then employed SVM for emotion prediction. Each component was trained independently, producing separate emotion predictions, which were subsequently combined using a weighted fusion strategy. Recently, Zheng et al. [70] developed the POSTER framework, a two-stream pyramid cross-fusion Transformer that processes both facial landmark and image features, achieving state-of-the-art results across three benchmark datasets.

3.3.2 Physiological Signals

Early approaches to emotion recognition using physiological signals primarily relied on handcrafted features and traditional machine learning methods. For instance, Setz et al. [71] proposed distinguishing stress from cognitive load in an office environment using EDA signals. From these signals, they extracted 16 time-domain features and employed six different classifiers, including linear discriminant analysis (LDA), SVM with linear, quadratic, polynomial, and RBF kernels, and the nearest class center (NCC) algorithm. Ragot et al. [72] compared emotion recognition performance between laboratory and wearable sensors. By recording EDA and cardiac activity from 19 participants using both the Biopac MP150 and Empatica E4 sensors, they extracted nine features and trained SVM classifiers to predict valence and arousal. Their results showed that wearable sensors are viable for non-intrusive emotion recognition, providing similar performance to laboratory sensors. In [73], the authors proposed training a Random Forest classifier using EDA signals and blood oxygen level data collected from 101 subjects experiencing various emotions. Hsu et al. [74] developed an ECG-based emotion recognition framework. They collected ECG signals while participants listened to music and extracted features through time-, frequency-, and nonlinear analyses. They employed a least squares SVM (LS-SVM), which achieved accuracies of 82.78% for valence, 72.91% for arousal, and 61.52% for emotion classification.

Despite these efforts, traditional methods faced challenges such as the need for domain-specific feature engineering and manual effort, which could lead to suboptimal performance. Deep learning approaches have addressed these limitations by automating feature

extraction and learning hierarchical data representations. Umematsu et al. [75] introduced a stress level recognition-based forecasting framework that leverages various data types, including physiological data, mobile phone usage, location information, and behavioral surveys collected over N days from 142 participants. Using Long Short-Term Memory (LSTM) neural networks, they achieved an accuracy of 83.6% in predicting next-day stress levels. Similarly, Awais et al. [76] employed a LSTM model to recognize emotions from physiological signals such as respiration, galvanic skin response (GSR), electrocardiogram (ECG), electromyogram (EMG), and temperature (TEMP). Dar et al. [77] proposed a hybrid approach combining Convolutional Neural Networks (CNNs) with LSTM networks for emotion recognition. More recently, Wierciński et al. [78] introduced GraphEmotionNet, a Graph Neural Network framework for emotion recognition using ECG, EEG, and GSR data from the AMIGOS dataset [79]. Their proposed model achieved accuracies of 69.71% for valence and 70.75% for arousal. In [80], the authors developed a Transformer-based self-supervised framework for emotion recognition from ECG signals. Their approach employs a CNN for feature extraction combined with a Transformer encoder, pre-trained using self-supervised learning on unlabeled ECG data and fine-tuned on the AMIGOS dataset for emotion recognition.

3.3.3 Multimodal Learning

Multimodal approaches to emotion recognition aim to combine data from different modalities to leverage the strengths of each to improve accuracy and robustness in classifying emotions. By combining behavioral cues (e.g., facial expressions, body movements) with physiological signals (e.g., heart rate, galvanic skin response), these methods can capture both external manifestations and internal states. Combining these modalities allows for a more holistic understanding of emotions, as some emotional states may not be fully represented by observable behavior alone.

Li et al. [81] introduced a multimodal facial expression recognition framework that integrates features from both EEG signals and facial landmarks at the input level. They extracted energy feature vectors from EEG signals using the discrete wavelet transform (DWT) and feature vectors derived from facial landmarks. These features are then concatenate and classified using a SVM. Soleymani et al. [25] developed a multimodal framework combining EEG, pupillary response, and gaze distance features to classify arousal (calm, moderate, high) and valence (unpleasant, neutral, pleasant) levels using SVM classifiers. Their study revealed that decision-level fusion outperformed both feature-level fusion and the use of individual modalities for both arousal and valence classifications. Saffaryazdi et al. [82] explored the fusion of facial micro-expressions, EEG signals, GSR, and photoplethysmography (PPG) for arousal and valence recognition. They compared individual modalities with multimodal fusion using SVM, Random Forest, K-Nearest Neighbors (KNN), and LSTM classifiers on the DEAP dataset [83] and another dataset collected by the authors. The multimodal fusion was achieved through either a voting scheme or weighted fusion. For facial micro-expressions, they employed 3D CNNs to process image sequences, while features from physiological signals were handled by either LSTM networks or traditional machine learning models. The fusion of all modalities using LSTM demonstrated superior

accuracy and F-score compared to the use of individual modalities across both datasets.

3.3.4 Discussion

Emotion recognition remains a challenging task due to the complex and dynamic nature of human emotions. Unimodal approaches relying on either behavioral or physiological signals offer valuable insights but are often limited by the inherent variability in human emotion expression. For example, facial expressions can be consciously masked or altered, while physiological signals can be influenced by non-emotional factors such as physical activity or health conditions.

Multimodal approaches offer a promising solution by integrating multiple data sources, capturing both external behaviors and internal physiological states. However, these approaches require sophisticated algorithms capable of aligning and fusing diverse data streams. Advances in deep learning and fusion techniques continue to push the boundaries of emotion recognition, making real-time, accurate emotion monitoring increasingly feasible across a variety of applications, from healthcare to human-computer interaction.

Given the importance of considering multiple modalities in emotion expression, the subsequent chapters will propose leveraging both behavioral and physiological data within our proposed multimodal frameworks for the tasks of stress detection and emotion recognition.

Chapter 4

Multimodal Transformer for Stress Detection

Contents

This chapter outlines our contributions to the field of automatic stress detection using physiological data. We introduce a novel multimodal Transformer framework that integrates various multimodal fusion techniques, including early fusion, intermediate fusion, and late fusion, to effectively integrate physiological signals from two sensors, treating each set of signals as a distinct modality. We establish both unimodal and multimodal benchmarks for automatic stress detection using the WESAD dataset. Additionally, we extend these benchmarks to the task of affect detection, further validating the effectiveness of our proposed multimodal framework.

In Section 4.1, we introduce the concept of stress, discuss its impact on health, and highlight the importance of automated stress detection systems. Section 4.2 provides a comprehensive review of the existing literature on automatic stress detection, establishing the necessary background for our proposed approach. In Section 4.3 we present our proposed framework. Following this, Section 4.4 presents the WESAD dataset, outlines the preprocessing steps applied to the physiological signals, and presents the results of our unimodal and multimodal stress and affect detection experiments. Section 4.5 discusses our findings, examining their implications, limitations, and potential avenues for future research. Lastly, Section 4.6 concludes the chapter by summarizing its key contributions.

4.1 Introduction

Stress has become a global epidemic and a significant concern, profoundly affecting individual lives and society as a whole. Various factors contribute to stress, including work-related pressures, financial difficulties, relationship issues, and social challenges [84]. Stress is a multifaceted response that involves physical, mental, and emotional reactions to stimuli that disrupt the typical state of balance. In response to stress, the body triggers the fight or flight response, a concept introduced by Walter Cannon [85]. This primal physiological reaction prepares the individual to either confront or escape the perceived threat.

While the fight-or-flight response is an evolutionary mechanism designed for survival, modern stressors are often more psychological than physical. Consequently, this response can become maladaptive in contemporary contexts. Stress can be broadly classified by duration into two distinct types: (1) acute stress, which is a brief, intense reaction to an immediate or unusual event or threat, and (2) chronic stress, which is a sustained, long-term physiological and psychological response to ongoing stressors or unresolved challenges.

In the context of medical education and practice, medical professionals, including students and doctors, often face high levels of stress due to the demanding nature of their education and profession. Stress can impact their overall performance and health. High levels of stress can impair cognitive functions such as memory, attention, and decision-making. For medical professionals, this impairment can lead to errors in judgment and decreased quality of care. Furthermore, chronic stress can increase susceptibility to certain types of cancer [86], slow wound healing [87], and increase vulnerability to infections [88].

Simulation practice, an integral part of medical training, provides a controlled yet realistic setting to replicate high-stress clinical scenarios. These simulations are designed to

prepare future healthcare professionals for real-world medical emergencies and decision-making under pressure. Monitoring stress during these practices can be highly beneficial for understanding how individuals respond to pressure. Additionally, it can provide valuable feedback on how stress levels impact performance during simulation sessions, enabling the development of personalized coping strategies to manage stress more effectively. These monitoring tools will facilitate targeted interventions and continuous improvements in medical training.

The physiological basis of stress involves a cascade of hormonal reactions initiated by the sympathetic nervous system (SNS). This response triggers the release of hormones such as ACTH, cortisol, and adrenaline, which impact various physiological parameters such as blood pressure, heart rate, and skin temperature, among others [89]. Monitoring these physiological parameters is essential for developing effective stress detection systems.

Machine learning and deep learning, have shown great promise for automatic stress detection using physiological data [90, 91]. In recent years, the Transformer [2] has revolutionized tasks in natural language processing (NLP) and demonstrated state-of-the-art performance across various tasks involving the processing of sequential data. These tasks include time series forecasting [12, 92, 93], time series classification [94, 95, 96], and anomaly detection in time series [97, 98, 99]. Furthermore Transformers has shown promising performance in multimodal learning for tasks such as emotion recognition [100], image classification [101], and action recognition [102], and more. Given its proficiency in handling sequential data, the Transformer model emerges as a strong candidate for automatic stress detection through sequences of physiological signals.

Given that stress can manifest through various physiological parameters, and that different physiological sensors can capture distinct types of signals with varying physical properties, it can be beneficial to consider multimodal approaches for automatic stress detection systems. Combining data from different types of sensors can enhance accuracy and robustness, as each modality can provide complementary information that improves overall detection performance. For instance, wrist-based sensors can measure physiological signals such as heart rate and skin conductance, while chest-based sensors can capture physiological signals like respiratory rate and ECG. This comprehensive data collection can lead to a more complete understanding of the body's stress response. Additionally, different sensors have varying susceptibilities to noise and artifacts, and multimodal data can help mitigate these issues, resulting in more reliable stress detection. Deep learning models benefit from the increased feature set provided by multimodal data, enhancing their ability to discriminate between stress and non-stress states. However, multimodal learning presents multiple challenges, including multimodal data fusion, synchronization, increased complexity and cost.

Motivated by the success of Transformer models for sequential data processing and multimodal learning, we propose a novel multimodal Transformer framework for automatic stress detection. Our framework incorporates different fusion strategies: early, intermediate, and late fusion, allowing for a comprehensive integration of physiological signals from two distinct types of sensors: wrist-based and chest-based. Through extensive experiments conducted on the WESAD dataset [90], we demonstrate that our approach surpasses all existing state-of-the-art methods for stress detection. To further validate the effectiveness

of our framework, we also conduct experiments on the task of affect detection.

The contributions of this work are three-fold and can be summarized as follows:

1. We establish benchmarks for both unimodal and multimodal stress and affect detection using physiological signals from two different sensors.

2. We propose a multimodal framework that integrates physiological signals from wrist and chest sensors, treating each set of signals as a separate input modality.

3. Our proposed approach achieves state-of-the-art results on the WESAD dataset.

4.2 Related Work

In recent years, physiological data have gained prominence in automatic stress detection due to their direct, objective measurements of the body's response to stress. Unlike video or voice data, which require interpretation and can be influenced by external factors or an individual's ability to conceal emotions, physiological signals provide a straightforward and objective indicator of physiological arousal related to stress.

The WESAD dataset, introduced by Schmidt et al. [90], has established itself as a benchmark for research in developing methods for automatic stress detection using physiological data. They included a comprehensive benchmark using signals from a wrist and chest sensor devices, either individually or in combination, to train classical machine learning models. Subsequent studies have leveraged more advanced methods on the WESAD dataset to enhance the stress detection performance. For instance, Samyoun et al. [91] proposed a sensor translation mechanism using Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs), and Multi-Layer Perceptrons (MLPs) to translate wrist data into chest-based features, subsequently applying classical machine learning methods for stress detection. Huynh et al. [103] proposed an optimized deep neural network training scheme using neural architecture search based on CNNs. In [104], the authors introduced the H-CNN framework, which comprises two main branches: the first branch processes handcrafted features, while the second branch incorporates multiple convolutional blocks. The output features from each branch are then concatenated before the classification stage. Wu et al. [105] explored the use of symmetric positive definite (SPD) matrices for the efficient integration of physiological and motion signals. Their method effectively captured correlation information both within each modality and between the two. The study demonstrated substantial performance improvements when multiple modalities were employed, compared to the use of a single modality.

Other studies have explored the integration of other types of modalities for automatic stress detection. Aigrain et al. [106] trained a SVM classifier on body movement features, facial expressions, and physiological data to identify stress. Mou et al. [107] proposed using attention mechanisms to fuse features from eye data, vehicle data, and environmental data for detecting driver stress.

In this chapter, we will present, to the best of our knowledge, the first application of

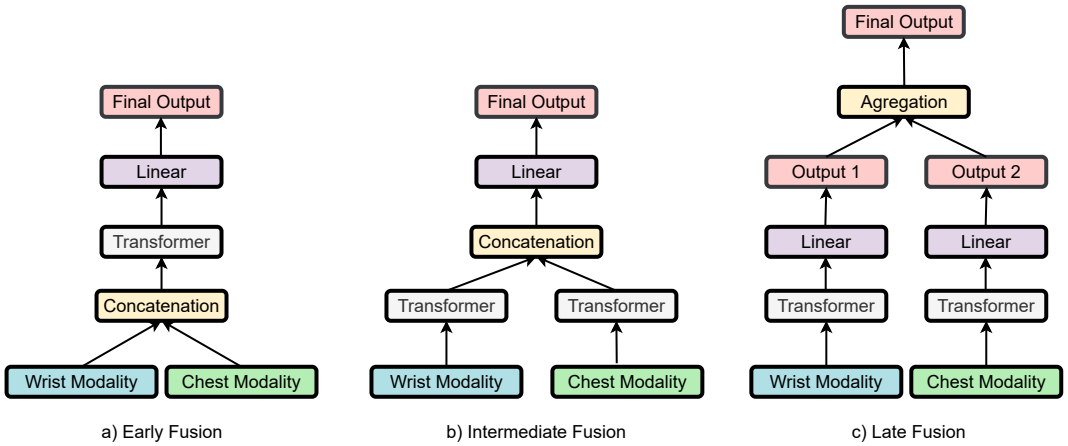


Figure 4.1: Illustration of the three main multimodal fusion strategies using Transformer encoders, namely early fusion (a), intermediate fusion (b), and late fusion (c).

multimodal Transformers for automatic stress detection using physiological data. We propose treating each set of physiological signals from wrist-based and chest-based sensors as distinct modalities. We employ a multimodal Transformer framework, incorporating three fusion strategies: early, intermediate, and late fusion.

4.3 Proposed Approach

The present section outlines the different multimodal fusion techniques that we have employed for stress detection using the Transformer architecture. These multimodal fusion techniques include early fusion, intermediate fusion, and late fusion, which are illustrated in Figure 4.1.

4.3.1 Multimodal Transformer

Early Fusion

Regarding the early fusion strategy, the raw physiological signals from the wrist and chest sensors are concatenated at the input level. This combined input is then fed into a single Transformer encoder that learns the joint representation of the multimodal data. Mathematically, if x_w and x_c represent the input signals from the wrist and chest sensors, respectively, the early fusion input x_{early} can be expressed as:

$$x_{early} = [x_w; x_c]$$

where ";" denotes the concatenation operation between both input sequences at the feature level. Next, the Transformer encoder processes the x_{early} sequence to produce the following representation:

$$z_{early} = \text{Transformer}(x_{early})$$

Intermediate Fusion

Intermediate fusion involves processing each modality independently through separate Transformer encoders and then combining their outputs at an intermediate layer. Let z_w and z_c denote the outputs of the wrist and chest Transformer encoders, respectively.

$$z_w = \text{Transformer}_w(x_w)$$
$$z_c = \text{Transformer}_c(x_c)$$

Subsequently, these outputs are fused using a concatenation operation followed by a fully connected layer to produce the combined representation z_{inter}:

$$z_{inter} = [z_w; z_c]$$

Late Fusion

In the late fusion strategy, the wrist and chest signals are processed separately through their respective Transformer encoders:

$$z_w = \text{Transformer}_w(x_w)$$
$$z_c = \text{Transformer}_c(x_c)$$

Then, each fused representation is projected using a different fully connected layer:

$$y_w = fc_w(z_w)$$
$$y_c = fc_c(z_c)$$

Mathematically, y_w and y_c are the logits from the wrist and chest modalities, respectively. Next, the outputs from these two fully connected layers are combined at the logit level. The combined logit $\bar{y}$ is calculated as:

$$\bar{y} = \frac{y_w + y_c}{2}$$

For the binary stress detection task, the final prediction $\hat{y}_{late}$ is obtained by applying the sigmoid function to $\bar{y}$:

$$\hat{y}_{late} = \sigma(\bar{y}) = \sigma\left(\frac{y_w + y_c}{2}\right)$$

4.3.2 Stress Classifier

For the early and intermediate fusion strategies, the stress classifier layer processed the fused representation, either z_{early} or z_{inter} to output the stress probability prediction. This classification head typically consists of a fully connected layer followed by a sigmoid activation function to produce the probability that the multimodal data sequence is associated to the stress class.

$$\hat{y} = \sigma(fc(z_{fusion}))$$

where σ is the sigmoid activation function, fc represents the fully connected layer and z_{fusion} denotes the fused representation from one of the two fusion strategies (early or intermediate), and $\hat{y}$ is the predicted stress probability.

4.4 Experimental Results

In this section, we will begin by providing a comprehensive overview of the WESAD dataset used in our experiments. Next, we will detail the different preprocessing steps applied to the raw physiological signals. Finally, we will present and analyze the results of our experiments on stress and affect detection tasks.

4.4.1 Dataset

The WESAD (Wearable Stress and Affect Detection) dataset is a comprehensive multimodal dataset widely used in research on automatic stress and affect detection using physiological data. It includes physiological and motion data from 15 healthy subjects (13 male, 2 female, aged 25-35), collected using two sensor devices: the RespiBAN Professional and the Empatica E4 wristband.

The Empatica E4 [108] records electrodermal activity (EDA), blood volume pulse (BVP), body temperature (TEMP), and three-axis acceleration (ACC) at frequencies of 4 Hz for EDA and TEMP, 64 Hz for BVP, and 32 Hz for ACC. An example of these signals for a given subject over an entire data collection period is shown in Figure 4.2. The RespiBAN device [109] measures electrocardiogram (ECG), electromyography (EMG), respiration (RESP), skin temperature (TEMP), EDA, and three-axis acceleration (ACC), all sampled at 700 Hz.

The experimental protocol of the WESAD dataset includes several phases designed to evoke different emotional responses: a baseline rest period, a stress test via the Trier Social Stress Test (TSST) [110], an amusement phase with the viewing of amusing videos, and a meditation phase for relaxation. Each session is annotated with labels corresponding to these phases (e.g., baseline, stress, amusement, meditation).

In alignment with previous studies [90, 91, 111, 103, 105, 112], we formulate a binary stress detection task (stress vs. no-stress) by combining the neutral and amusing stimulus

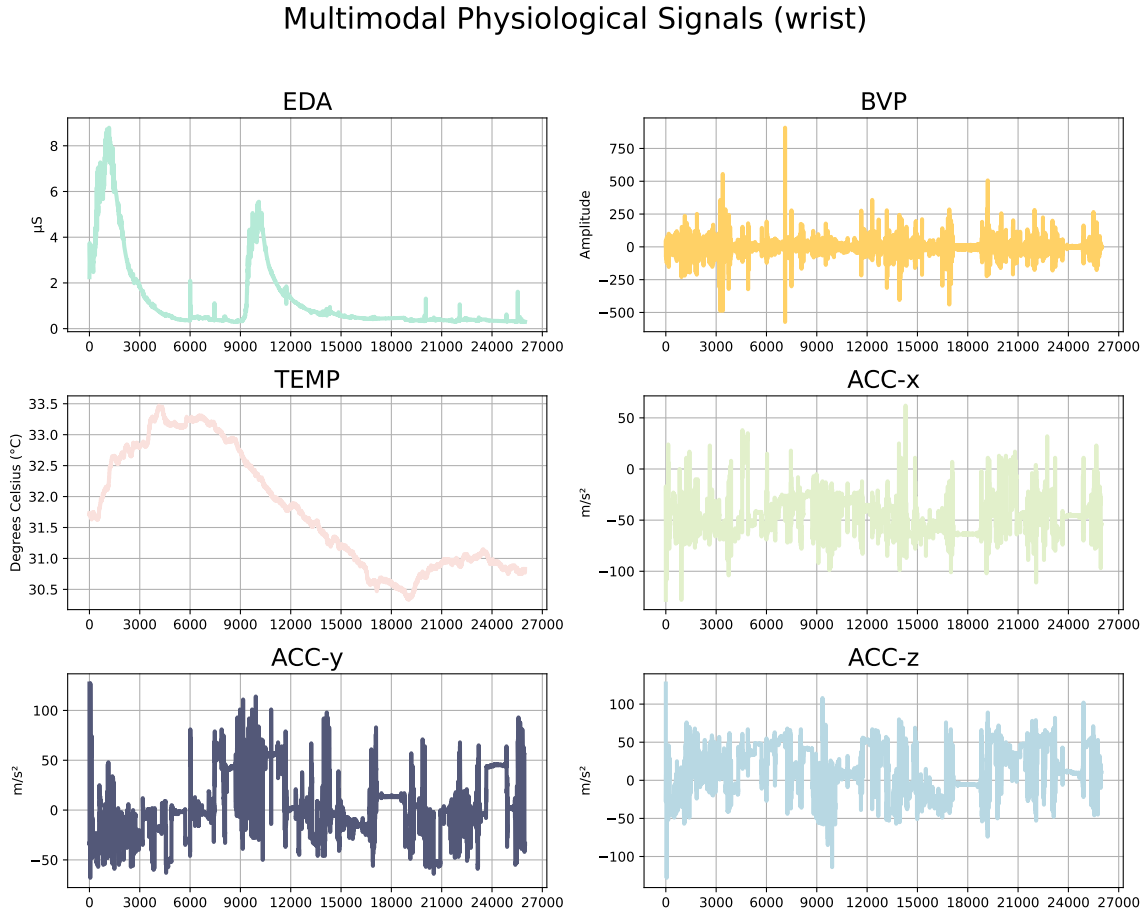


Figure 4.2: Plot of the electrodermal activity (EDA), blood volume pulse (BVP), body temperature (TEMP), and three-axis acceleration (ACC) data for a subject from the WESAD dataset. Data collected using the Empatica E4 wristband spans from the start to the finish of the recording session, with the x-axis representing the time steps.

sequences as the "no-stress" class.

4.4.2 Preprocessing

In the following, we will describe the different preprocessing steps applied to physiological signals collected from the wrist and chest sensors. The overall pipeline, which include the Multimodal Transformer framework (MMT), is illustred in Figure 4.3.

Filtering

We applied a low-pass filter to all the physiological signals from both sensors to reduce noise and preserve the frequencies of interest. Next, we downsampled the signals from the Empatica E4 wrist sensor to 4 Hz, matching the smallest frequency present in all the

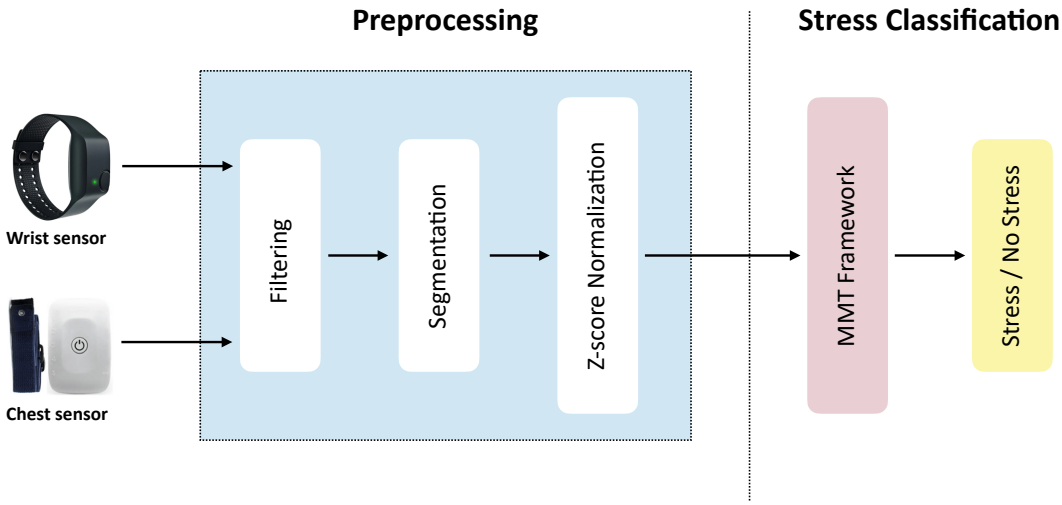


Figure 4.3: A flowchart illustrating each preprocessing step alongside the proposed stress classification framework.

signals from this device, which comes from the EDA. For the chest sensor, we retained the sampling frequency of 700 Hz for all physiological signals. Regarding our MMT framework, we downsampled all signals from both sensors to 4 Hz, ensuring consistency with the lowest frequency present across all signals from both sensors, which is EDA from the wrist sensor.

Segmentation

Next, we segmented all physiological signals into 60-second sliding windows without any overlap between successive windows.

Normalization

To ensure that the physiological signals are on a comparable scale, we applied Z-score normalization to all signals. Z-score normalization involves transforming the data so that it has a mean of zero and a standard deviation of one. This is achieved by subtracting the mean of each signal from its values and then dividing the result by the signal's standard deviation. The formula for Z-score normalization for a given physiological signal is:

$$Z = \frac{X - \mu}{\sigma}$$

where X is the original signal value, μ is the mean of the signal, and σ is the standard deviation of the signal.

4.4.3 Implementation Details

To determine the optimal hyperparameters for our proposed framework, we employed a grid-search strategy, exploring the following hyperparameters with their respective values:

- Dimension of the linear projection layer: 256 and 512
- Number of multi-head attention: 4 and 8
- Number of Transformer encoder layers: 1 and 2
- Dropout rate in the Transformer encoder: 0.0 and 0.2

The batch size was fixed at 32, the maximum number of epochs at 150 with an early stopping patience criterion set at 70 epochs. The learning rate was varying between 10^{-3} and 10^{-5}. All models were trained using the Adam optimizer [113], with exponential decay rates for the first and second moment estimates set at 0.9 and 0.999, respectively. The entire framework was implemented using the PyTorch library [114].

4.4.4 Evaluation Framework

Following prior works on the WESAD dataset [90, 105, 91, 111, 103, 112], we employ the Leave-One-Subject-Out Cross Validation (LOSO-CV) evaluation procedure to validate our models. Given the 15 subjects in our dataset, this procedure involves training our models on 14 subjects and testing on the remaining subject, repeating this process for all 15 subjects. For evaluation metrics, we used accuracy and weighted average F1-score, and we report both the mean and standard deviation of these metrics across the 15 folds.

4.4.5 Results

In the following, we will present and analyse our results on both unimodal and multimodal stress and affect detection. For the affect detection task, the goal is to classify physiological sequences into one of the three following class: baseline, stress, and amusement.

Stress Detection

Unimodal: The Transformer model outperformed all other methods in terms of accuracy and F1-score when using wrist-based physiological signals, surpassing the best-performing model [105] by approximately 1.61% and 2.08%, respectively, as we can see in Table 4.1. Specifically, the Transformer model achieved an accuracy of 96.26% and an F1-score of 96.07%. Similarly, for chest-based data, the Transformer model achieved the best results with an accuracy of 97.20% and an F1-score of 97.20%, outperforming the best state-of-the-art by 0.51% and 0.59% in terms of accuracy and F1-score, respectively.

It is noteworthy that deep learning-based methods [112, 103, 105, 111] consistently outperformed machine learning-based methods [90, 91] for both types of sensors by a signif-

Table 4.1: Unimodal stress detection: comparison with state-of-the-art methods.

Methodes	Wrist		Chest	
	Acc	F1 score	Acc	F1 score
Schmidt et al.[90]	87.12	84.11	92.83	91.07
Samyoun et al. [91]	89.90	87.60	91.10	90.20
Gil-Martin et al. [111]	92.70	92.55	93.10	93.01
Huynh et al. [103]	93.14	-	-	-
Wu et al. [105]	94.65	93.99	95.54	94.76
Lai et al. [112]	94.16	93.62	96.69	96.61
Transformer	**96.26 ± 5.63**	**96.07 ± 5.94**	**97.20 ± 4.44**	**97.20 ± 4.32**

icant margin. This trend highlights the superiority of deep learning models for accurate stress detection, primarily due to their ability to learn directly from raw data without the need for manually engineered features. In contrast, machine learning methods [90, 91] rely heavily on feature engineering, which can limit their performance.

Additionally, the Transformer model demonstrated better performance using physiological signals from the chest sensor compared to those from the wrist sensor. However, it is important to note that the difference in performance, while present, is not very large, indicating that wrist-based sensors, despite being less intrusive and more convenient, still provide highly valuable data for accurate stress detection.

Multimodal: We present the performance of our MMT framework in Table 4.2. Our three proposed multimodal architectures, MMT-early, MMT-inter, and MMT-late outperformed all other state-of-the-art models by a large margin. Specifically, MMT-inter and MMT-late demonstrated superior performance, achieving the highest accuracy and F1-scores. MMT-inter showed an improvement of 1.31% in accuracy and 1.33% in F1-score compared to the best-performing model [112]. MMT-inter achieved an accuracy of 99.06% and an F1-score of 99.07%.

These results confirm our hypothesis that the use of multimodal models is appropriate for processing groups of signals from different sensors. The superior performance of the MMT-inter model indicates that the intermediate fusion strategy, which combines features from different sensors at a mid-level stage, is particularly effective. This suggests that retaining a certain level of independence in the early stages of processing while merging the information in the intermediate stage allows for better extraction and integration of relevant features from both wrist and chest sensors. Similarly, the results of the MMT-late model, which employs a multimodal Transformer with a late fusion strategy, highlight the efficacy of processing each sensor's data independently until the final stages. This method can also lead to robust feature extraction and integration. Both approaches underscore the versatility and effectiveness of multimodal frameworks in leveraging heterogeneous phys-

Table 4.2: Multimodal stress detection: comparison with state-of-the-art methods.

Methodes	Wrist + Chest	
	Acc	F1 score
Schmidt et al.[90]	92.28	90.74
Samyoun et al. [91]	94.70	93.40
Gil-Martin et al. [111]	96.62	96.63
Wu et al. [105]	96.88	96.44
Lai et al. [112]	97.75	97.74
MMT-early (ours)	98.88 ± 2.24	98.90 ± 2.19
MMT-inter (ours)	**99.06 ± 1.67**	**99.07 ± 1.64**
MMT-late (ours)	**99.06 ± 1.69**	99.05 ± 1.73

iological data from multiple sensors for improved performance.

Table 4.5 presents the highest accuracy and F1-score achieved for each subject in the test set using the MMT-inter architecture. Notably, for the majority of subjects (S2, S4, S5, S6, S7, S8, S9, S10, S13, S15, S17), the MMT-inter model achieved a perfect score of 100.00% in both accuracy and F1-score for stress detection. A few subjects (S3, S11, S14, S16) exhibited slightly lower scores, but they still maintained high accuracy and F1-scores, all exceeding 94%.

Affect Detection

Unimodal: For the affect detection task, the Transformer model significantly outperformed all other methods, whether using wrist-based or chest-based physiological data, as shown in Table 4.3. Specifically, for the wrist-based physiological data, the Transformer achieved an accuracy of 78.15% and a F1-score of 74.53%, surpassing the best performing state-of-the-art model by 2.94% and 10.41% in terms of accuracy and F1-score, respectively. For the chest-based sensor, the Transformer reached an accuracy of 83.92% and a F1-score of 78.53%, exceeding the best performing state-of-the-art model by 7.42% in accuracy and 6.04% in F1-score.

It is worth noticing that the gap between the accuracy and F1-score is significant for most state-of-the-art comparison models, indicating that while these models perform well overall, they struggle with some classes. This contrasts with the Transformer model, which shows a smaller gap between accuracy and F1-score, suggesting a more balanced performance across all classes.

Additionally, similar to the stress detection task, the Transformer performs better when using chest-based physiological data compared to wrist-based data. However, for the af-

Table 4.3: Unimodal affect detection: comparison with state-of-the-art methods.

	Wrist		Chest	
Methodes	Acc	F1 score	Acc	F1 score
kNN [90]	45.54	37.20	46.18	38.39
Decision Tree [90]	53.98	43.62	57.68	53.06
Random Forest [90]	74.85	62.86	68.76	60.80
AdaBoost [90]	75.21	64.12	74.74	64.89
LDA [90]	70.74	63.24	76.50	72.49
Transformer	**78.15 ± 16.11**	**74.53 ± 16.61**	**83.92 ± 6.13**	**78.53 ± 8.01**

Table 4.4: Multimodal affect detection: comparison with state-of-the-art methods.

	Wrist + Chest	
Methodes	Acc	F1 score
kNN [90]	56.14	48.70
Decision Tree [90]	65.56	58.05
Random Forest [90]	74.97	64.08
AdaBoost [90]	79.57	68.85
LDA [90]	75.80	71.56
MMT-early (ours)	88.28 ± 6.16	**85.33 ± 8.69**
MMT-inter (ours)	88.03 ± 4.96	84.94 ± 7.51
MMT-late (ours)	**88.41 ± 6.04**	84.79 ± 9.07

fect detection task, the difference in performance for both evaluation metrics is more pronounced than for the stress detection task between both sensors, with differences of 5.77% and 4.00% in accuracy and F1-score, in favor of the chest-based sensor.

Multimodal: As shown in Table 4.4, our proposed multimodal methods—MMT-early, MMT-inter, and MMT-late-far exceeds all state-of-the-art methods. Notably, MMT-late achieved the highest accuracy, surpassing the best existing method by 8.84%. On the other hand, MMT-early delivered the best performance in terms of weighted F1-score, with a score of 85.33%, representing an improvement of 16.48% compared to the best state-of-the-art method.

Table 4.6 reports the best accuracy and F1-score for each subject in the test set when employing MMT-late. The highest performance is observed for subject S15 with accuracy

and F1-score of 97.22% and 97.26% respectively. Several subjects (S6, S9, S17) also exhibit high scores, particularly above 90%. However, some subjects show lower results, with the lowest performance observed in subject S6, who has an accuracy of 80.56% and an F1-score of 74.06%.

Subjects	Acc	F1
S2	100.00	100.00
S3	94.29	94.08
S4	100.00	100.00
S5	100.00	100.00
S6	100.00	100.00
S7	100.00	100.00
S8	100.00	100.00
S9	100.00	100.00
S10	100.00	100.00
S11	97.22	97.18
S13	100.00	100.00
S14	97.22	97.18
S15	100.00	100.00
S16	97.22	97.25
S17	100.00	100.00
Average	**99.06**	**99.05**

Table 4.5: Best accuracy and F1-score for each subject in the test set using MMT-inter.

Subjects	Acc	F1
S2	88.57	86.31
S3	88.57	88.50
S4	82.86	75.45
S5	82.86	76.26
S6	97.14	97.08
S7	82.86	75.45
S8	88.89	86.64
S9	97.14	97.08
S10	83.78	76.78
S11	91.67	90.78
S13	83.33	76.14
S14	80.56	74.06
S15	97.22	97.26
S16	83.33	76.79
S17	97.30	97.23
Average	**88.41**	**84.79**

Table 4.6: Best accuracy and F1-score for each subject in the test set using MMT-late.

4.5 Discussion

Our findings regarding our proposed framework for automatic stress detection have significant implications for both affective computing research and practical health monitoring applications. In this section, we will discuss the impact of our results, the potential limitations of our study, and future directions for research in this field.

4.5.1 Implications of Findings

One of the key contributions of this research is the establishment of a benchmark for stress and affect detection using physiological signals collected from both wrist-based and chest-based sensors. Our approach achieves state-of-the-art performance on the WESAD dataset,

demonstrating its effectiveness in stress detection. The proposed dual-modality approach leverages effectively the strengths of each sensor type, resulting in a better understanding of stress responses. Wrist-based sensors, for instance, provide continuous monitoring of heart rate and skin conductance, which are critical indicators of stress. Chest-based sensors, on the other hand, offer precise measurements of respiratory rate and ECG. The combination of these signals enables the multimodal Transformers to capture a comprehensive profile of the body's stress response, thus enhancing the accuracy and reliability of stress detection.

The practical implications of our research are vast. The proposed multimodal Transformer framework can be implemented in wearable devices and health monitoring systems to provide continuous, real-time stress monitoring. Such systems will be able to offer timely interventions, personalized stress management strategies, and early warnings to prevent stress-related health issues. This is particularly relevant in high-stress professions, healthcare, and personal wellness, where effective stress management can significantly improve quality of life and productivity.

4.5.2 Limitations

Our experiments were conducted on the WESAD dataset, which, although comprehensive, may not fully represent the diversity of stress responses across different populations. Future research should include diverse and larger datasets to validate the generalizability of our framework. Cross-dataset validation is essential for ensuring that the model performs consistently across various demographic, physiological profiles, and data collection protocols.

Furthermore, the definition and labeling of stress can vary significantly between datasets, affecting the consistency and comparability of stress detection models. Labeling strategies range from self-reported assessments, where participants rate their stress levels using questionnaires, to physiological markers such as EDA, task-induced stress (as used in the WESAD dataset), and expert-defined labels based on psychological evaluations. Due to the complexity and subjectivity of stress, criteria from one dataset may not be directly applicable to another, creating challenges for model training and evaluation. This highlights the need for standardized definitions and protocols for stress assessment.

Additionally, the performance of our framework relies on the availability and accuracy of physiological sensors. Variability in sensor quality and placement can affect the reliability of the data, and consequently, the performance of the stress detection model. This variability in sensor input may degrade the model's performance, leading to unreliable predictions and impacting its overall effectiveness, especially in real-world applications where sensor quality cannot always be guaranteed.

One important limitation is the lack of model prediction interpretability. Although our multimodal Transformer-based framework demonstrates high accuracy in detecting stress, the decision-making process of the model remains a "black box." Transformers, and deep learning models in general, are inherently complex, and it is challenging to explain how the model arrives at a specific prediction. This limitation can reduce the trust of healthcare professionals, where understanding the rationale behind a prediction is essential for users

and healthcare providers. Without interpretability, it is difficult to identify which specific features or sensor signals (e.g., heart rate, respiratory rate, skin conductance) are driving the model's decision, potentially limiting the model's utility in clinical settings.

Lastly, the computational demands of Transformer models, especially in a multimodal setup, present another limitation. These models require substantial processing power, which can be problematic for real-time stress detection applications on wearable devices with limited computational resources and battery life. The need for high computational capacity may restrict the framework's deployment in everyday wearable technologies, reducing its accessibility and effectiveness in real-world, continuous stress monitoring scenarios.

4.5.3 Future Directions

Several avenues for future research can build upon our findings. One promising direction is the inclusion of other sensor modalities such as facial expression analysis, facial landmarks or even gaze tracking. These could complement physiological signals and provide richer data for stress detection models, potentially improving the accuracy of stress classification by leveraging more nuanced markers of emotional states.

Another important area of focus is the development of personalized models. Stress responses vary widely across individuals due to physiological differences, personal health conditions, psychological factors, and environmental influences. To address this variability, future research should investigate the development of models specifically tailored to individual users.

To address the limitations of dataset variability, future research should explore the development of models capable of cross-dataset generalization. This would involve training and testing on a diverse range of datasets to improve robustness across different populations, sensor types, and data collection protocols.

Enhancing the explainability of the model's predictions will be crucial for gaining user trust and acceptance, especially in health monitoring systems. Future research should explore explainable AI (XAI) methods that can highlight which physiological features or sensor signals are most influential in determining stress levels. Providing users and clinicians with interpretable feedback will foster trust, improve understanding, and support informed decision-making.

Given the computational complexity of multimodal Transformers, future work should focus on optimizing these models for real-time stress monitoring in wearable devices. Techniques such as model compression, and pruning can be explored to reduce computational overhead while maintaining performance. This is essential for deploying these systems in real-world applications where power consumption and processing limitations are critical constraints.

4.6 Conclusion

In this chapter, we introduced a novel multimodal Transformer framework for the tasks of stress and affect detection, using physiological signals from both wrist-based and chest-based sensors. Our approach leverages multiple multimodal fusion strategies, including early, intermediate, and late fusion strategies. The proposed framework achieved state-of-the-art performance on both tasks using the WESAD dataset.

The success of our framework can be attributed to two key factors. First, the unimodal Transformer encoder, which achieved state-of-the-art results in stress and affect detection when using physiological signals from each sensor individually. Second, treating the physiological signals from each sensor as distinct modalities enabled the model to integrate complementary information from both wrist-based and chest-based sensors in an effective way, providing a more comprehensive understanding of the body's stress response.

Our work contributes to the ongoing development of more accurate and reliable stress detection systems, particularly in contexts where multimodal data can provide a more comprehensive understanding of physiological responses. While our approach shows promise, it also underscores the broader challenges associated with validating automatic stress detection systems. These challenges include the variability in physiological responses across different populations, the need for robust and diverse datasets, and the difficulty in ensuring that these systems can generalize effectively to real-world scenarios. Addressing these challenges is crucial for advancing the field and ensuring that stress detection technologies can be widely and reliably implemented.

Chapter 5

MMGT: Multimodal Graph-based Transformer for Pain Detection

Contents

This chapter introduces technical contributions related to multimodal data fusion in the context of deep learning for automatic pain detection and, more broadly, affective computing. We propose using a Graph Neural Network to efficiently fuse hierarchical representations derived from multiple data modalities, including facial landmarks, facial action units, and physiological signals, which are extracted using the Transformer architecture.

In Section 5.1, we discuss the nature of pain, and the necessity for pain detection systems. Section 5.2 reviews pertinent studies on both unimodal and multimodal pain detection, providing a comprehensive background for our proposed methodology. The details of our multimodal framework are presented comprehensively in Section 5.3. Following this, we present in Section 5.4 the public dataset used in our study, detail the data preprocessing pipeline, and discuss the results of our unimodal and multimodal pain detection experiments. Finally, in Section 5.5 we discussed the implications, limitations, and future directions of our work, and in Section 5.6 we briefly summarize our work.

5.1 Introduction

Pain is a complex, multifaceted experience that significantly impacts our well-being, encompassing both physical and psychological aspects. It serves as a sensory and emotional response to actual or potential tissue damage, acting as a crucial warning system to encourage protective actions. However, its subjective nature poses significant challenges in assessment and management, necessitating more objective, reliable, and efficient approaches to pain detection.

The emotional aspect of pain is a critical dimension that influences how individuals experience and cope with it. Pain often involves strong negative emotions such as distress, suffering, and discomfort, which can amplify its perception and make it more challenging to endure. Additionally, the anticipation of pain or fear of its recurrence can lead to anxiety, and past experiences of pain can leave emotional imprints that affect how new pain is perceived.

Understanding how pain is expressed through various physiological and behavioral responses is essential in developing automatic pain detection systems. Pain expression involves physiological changes such as heart rate, blood pressure, respiratory rate, skin conductance, pupil dilation, skin temperature, endocrine responses, and brain activity. Additionally, behavioral responses, including reflexive actions and pain behaviors like grimacing or verbal complaints, play key roles in communicating pain. These expressions provide essential cues for designing comprehensive pain detection models.

Given that pain is a multimodal experience, designing an effective pain detection model necessitates the incorporation of multiple modalities, such as physiological signals and facial expressions. Each modality has distinct statistical properties, and examining their interrelationships can yield valuable insights for improving pain detection. Integrating multiple data modalities offers several benefits: increased accuracy through capturing various pain aspects and enhanced robustness against the limitations or noise of single modalities.

Recent studies have demonstrated that multimodal interactions between the intermediate representations of deep neural networks can yield highly successful results across various applications. Notably, research indicates that leveraging these intermediate layers may be as effective, or even more advantageous, than relying solely on final-layer representations [115, 116, 117]. Building on these insights and inspired by the demonstrated success of Transformers and Graph Convolutional Neural Networks (GCNs) in multimodal tasks [118, 43], we propose to explore their potential for pain detection in multimodal settings.

In this work, we introduce a new multimodal fusion framework that leverages the capabilities of GCNs to explore interactions across various levels of modality-specific representations. Our proposed Multimodal Graph-based Transformer (MMGT) is built upon the intermediate Transformer representations of each modality. Specifically, we model these interactions through a graph structure where each node corresponds to a feature at a particular level within a modality. Nodes are connected both within a modality and across modalities.

The combination of the GCN and Transformer in a complementary setting provide a powerful framework for capturing both modality-specific and cross-modal relationships. The Transformer layers focus on modeling intra-modal relationships, while the GCN layers facilitate the exploration of interactions across different modalities. This synergy enables our model to better handle the heterogeneity of the input data and uncover complex patterns that would be difficult to capture with traditional fusion techniques for accurate pain detection.

To verify the effectiveness of our proposed approach, we conducted extensive experiments on the BP4D+ dataset [119]. Our MMGT model outperforms all existing multimodal state-of-the-art methods for the task of pain detection using combinations of two and three modalities, including physiological signals, facial action units, and facial landmarks (2D/3D/Thermal).

The contributions of this work are three-fold and can be summarized as follows:

1. The proposition of a new multimodal fusion framework that leverages a GCN to efficiently combine representations extracted by unimodal Transformer encoders from different modalities.

2. Demonstration of the complementarity between the modalities through benchmarking using single modalities and different combinations of two and three input modalities using our MMGT.

3. To the best of our knowledge, our MMGT is the first multimodal model trained on facial landmarks, facial action units, and physiological data.

5.2 Related Work

Automatic pain detection is a complex and multifaceted field that leverages various modalities to accurately assess and quantify pain levels in individuals. This section reviews the

existing body of work in pain detection, categorizing the research into unimodal and multimodal approaches.

5.2.1 Unimodal Pain Detection

Unimodal pain detection involves using a single source of information to identify and assess pain. This approach can be categorized into static and temporal methods. Static methods utilize individual data points to detect pain, while temporal methods analyze sequences of data over time to achieve the same goal.

Static Approaches

Vision-based modalities are highly favored in the design of pain detection systems due to their ability to capture facial expressions, which are crucial for communicating pain to others. The UNBC-McMaster Shoulder Pain dataset [120] has been a pioneering effort towards the development of pain detection methods from facial expressions. This dataset consists solely of facial images captured from 129 participants experiencing pain in one of their shoulders. The facial expressions of the participants were recorded using a digital camera as they underwent eight range-of-motion evaluations on their affected and unaffected shoulders. Early methods employing this dataset for pain detection relied on handcrafted features. For instance, Khan et al. [121] extracted both the pyramid histogram of orientation gradients (PHOG) for shape information and the pyramid local binary pattern (PLBP) for appearance information from the upper and lower facial regions, which were then combined and processed using traditional machine learning methods for pain detection. In [122], the authors proposed a method to recognize pain in images of faces with occlusions by simulating occlusion in the UNBC-McMaster pain dataset. They extracted the discrete cosine transform (DCT), local binary pattern (LBP), and histogram of oriented gradients (HOG) descriptors from a small window around the eye and subsequently employed a linear support vector machine (SVM) for pain detection. Florea et al. [123] introduced the Histogram of Topographical (HoT) features to characterize the face, alongside a novel transfer learning method to estimate pain intensity.

More recent developments in the field have seen a shift towards deep learning-based approaches, largely due to their impressive performance in extracting complex patterns from facial expressions. This capability is particularly valuable in pain detection, where expressions of pain can be subtle and nuanced.

Zamzmi et al. [124] proposed the Neonatal Convolutional Neural Network (N-CNN) for assessing neonatal pain. To validate their approach, they collected a dataset consisting of 31 neonates hospitalized and recorded in the Neonatal Intensive Care Unit (NICU). In addition, they experimented with transfer learning by fine-tuning a VGG-16 [69] and a ResNet-50 [67]. The authors in [125] fine-tuning a ResNeXt [126] model to detect pain in the masked faces of patients undergoing procedural sedation in the Interventional Radiology department of a hospital. El Morabit et al. [127] fine-tuned the data-efficient image transformers (DeiT) [128] architecture for pain detection. Yuan et al. [129] recently in-

troduced a pain assessment framework comprising three modules: an AU-guided module (AUG), a texture feature extraction module (TFE), and a pain assessment module (PA). The AUG module is responsible for detecting facial Action Units (AUs) from the unoccluded regions of the face. The detected AUs are then fed into a linear network that outputs fixed-size vectors. This module allows the network to learn the expression of pain features more effectively, enhancing residual feature learning in the presence of occlusion. The TFE module utilizes the Mediapipe [130] framework to identify facial landmarks and crop out three types of patches—prior-knowledge patches, a randomly explored patch, and a global feature patch—for extracting texture features through convolution modules. These texture features, along with AU information, are then integrated in the PA module to evaluate a patient's current pain status.

Temporal Approaches

Nevertheless, expressions of pain can be complex and subtle, often involving a sequence of facial movements rather than a single static expression. Analyzing sequences allows the detection system to capture these dynamic patterns. Szczapa et al. [131] proposed a framework for trajectory analysis on video sequences based on the computation of Gram matrices from 66 facial points and their velocities for estimating pain intensity. Several studies [132, 133, 134, 135] employed a combination of Convolutional Neural Network and Recurrent Neural Network for spatio-temporal features learning. Xu et al. [136] introduced the PET framework, which consists of an initial image encoding module featuring a ResNet-50 combined with an attention block for learning spatial features, and a Transformer module to learn temporal dependencies among video frames.

Several studies have employed physiological signals to recognize pain, offering the advantage of being more objective by directly measuring physiological changes associated with pain. Susam et al. [137] proposed the use of time-scale decomposition (TSD) to analyze the electrodermal activity (EDA) signal, which measures short- and long-term changes in time series data. They subsequently applied a linear Support Vector Machine (SVM) to the extracted features for pain detection. Chu et al. [138, 139] proposed in two successive studies, the used of EDA, ECG, and BVP for pain intensity estimation. Other studies explored the use of brain activity for pain recognition using EEG [140, 141], fMRI [142], and fNIRS [143]. However, there are several drawbacks to many contact-based sensors: (1) they can be sensitive to motion artifacts or other external interferences, such as electrical noise or environmental factors; (2) depending on the type of sensor and its placement, wearing physiological sensors can be uncomfortable and invasive; (3) physiological sensors often require specialized equipment that can be expensive to procure and maintain.

5.2.2 Multimodal Pain Detection

Pain can be expressed through multiple modalities, such as facial expressions and physiological signals, and body movements. For that reason, multimodal learning can greatly benefit automatic pain detection. Single-modality approaches in pain detection, such as using only facial expressions or physiological signals, often fall short in capturing the full

complexity of pain experiences. Each modality, while useful, has inherent limitations. For example, facial expressions can provide significant clues about pain but are subject to individual differences and cultural variations. Similarly, physiological signals offer objective data but can be influenced by factors unrelated to pain, such as stress or physical activity. By integrating multiple data sources, such as facial expressions, and physiological signals, multimodal methods can offer a more comprehensive and accurate assessment of pain. These approaches not only improves the accuracy and reliability of pain detection but also allows the strengths of one modality to offset the weaknesses of another.

Zhi et al. [144] proposed a multimodal framework for pain assessment that utilizes multiple branches to capture spatiotemporal features from facial expressions. This is achieved using raw facial video frames and optical flow images at different frame rates. These facial features are then fused with physiological features extracted from an LSTM, and the combined features are employed for pain assessment. Salekin et al. [145] introduced a multi-channel deep neural network framework for detecting pain in neonates. Initially, they employed YOLOv3 [146] for face and body detection. Subsequently, they used a VGG16 for feature extraction from the face, body, and their combination. The extracted features are concatenated into a single feature vector. To capture temporal correlations, the concatenated feature vectors from each frame are processed using an LSTM, which is then followed by a fully connected layer to perform pain detection. In a subsequent study, Salekin et al. [147] proposed incorporating neonatal crying sounds alongside facial and body information for pain detection. Their results demonstrated that the sound modality significantly outperformed facial and body modalities in detecting pain, underscoring the value of integrating all three modalities. They showed that their framework, which utilized all three modalities, surpassed unimodal methods that relied on any single modality. Hinduja et al. [148] proposed fusing physiological data and facial action units for pain detection. They trained a Random Forest classifier on these fused modalities and included cross- and gender-specific experiments. Their results showed that the fusion of both modalities outperformed unimodal approaches. For a comprehensive review of automatic pain detection, the reader may refer to [149, 150].

Previous studies on multimodal pain detection have primarily focused on integrating modalities through conventional fusion techniques, with limited exploration of advanced fusion methods that can capture intricate cross-modal relationships. Additionally, the potential of leveraging intermediate feature representations across modalities for pain detection has remained largely unexplored. In this chapter, we present a novel framework that addresses these gaps by introducing our proposed MMGT architecture. Our framework employs unimodal Transformer encoders to extract intermediate representations from each modality. These representations are then structured into a graph and processed through a GCN to facilitate efficient fusion and discovery of complex patterns across modalities.

To the best of our knowledge, this approach is the first to leverage physiological data, facial action units, and facial landmarks (2D, 3D, and Thermal) collectively for pain detection. Moreover, the combination of Transformer encoders and GCNs enables our framework to simultaneously capture both modality-specific features and cross-modal interactions, offering a more comprehensive understanding of the multimodal data involved in pain expression.

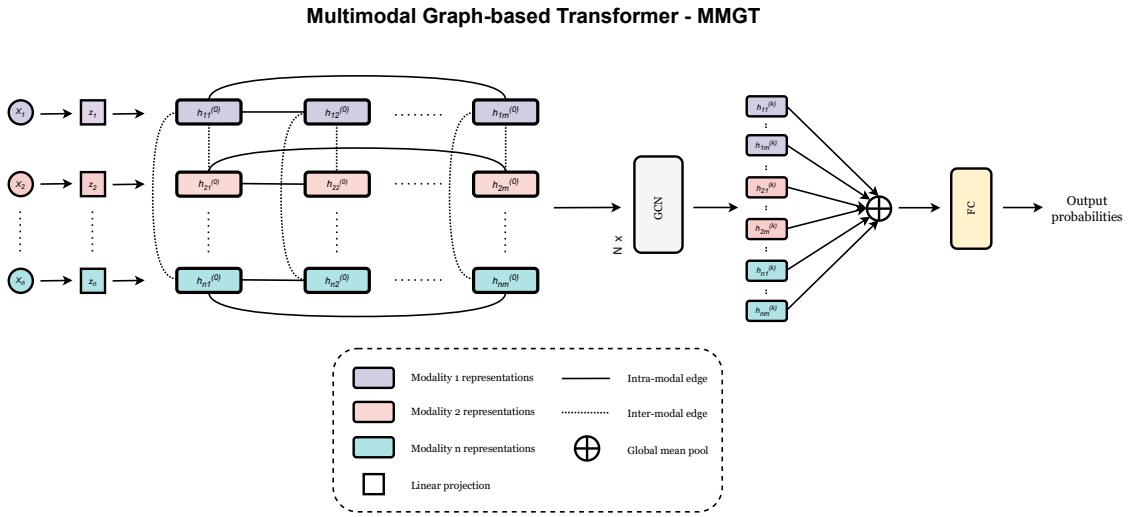


Figure 5.1: Illustration of our MMGT framework, which is composed of two main building blocks: Unimodal Transformer Encoders, and Multimodal Graph Convolutional Network.

5.3 Proposed Approach

This section introduces our MMGT framework. Figure 5.1 illustrates the proposed architecture, which consists of two primary components: (1) unimodal Transformer encoders that extract m intermediate feature representations for each modality; and (2) a GCN that efficiently fuses these extracted feature representations based on a graph representation that models the connections between the different representations across the different modalities. In the following, we provide a detailed review of each component.

5.3.1 Unimodal Transformer Encoders

The first stage of our framework involves linearly projecting the data from each input modality $x_1, .., x_n$ to embedding vectors, $z_1, \ldots, z_n$, each with a dimension of d_m. These projections are performed using the following learnable weight matrices $W_{x_1} \in \mathbb{R}^{d_{x_1} \times d_m}$, $\ldots$, $W_{x_n} \in \mathbb{R}^{d_{x_n} \times d_m}$, where $d_{x_1}, \ldots, d_{x_n}$ denote the dimensions of each respective input modality. Positional encodings are then added to each embedding vector to preserve the relative order within the sequences. Since the embeddings vectors and positional encodings share the same dimension, they are directly summed. Subsequently, these updated embedding vectors are fed into unimodal Transformer encoders, which extract a set of m intermediate feature representations for each input modality. For instance, for a given input modality x_i, the resulting representations are $h_{i1}, \ldots, h_{im}$.

5.3.2 Multimodal Fusion GCN

To effectively capture relationships among feature representations across different input modalities, we propose a graph-based approach where relationships are represented as con-

nections in a graph, and a GCN is used to model these interactions.

Graph Construction

Our proposed multimodal framework is based on the construction of a graph $\mathcal{G} = (V, E)$, where V represents the set of nodes, and E denotes the set of edges capturing relationships between these nodes. The graph is constructed as follows:

- **Nodes:** Each modality i is represented by m nodes, which are initialized using the previously extracted feature representations $h_{i1}, \ldots, h_{im}$. Given n input modalities, the notal number of nodes is $n \times m$.

- **Edges:** The connections among nodes are categorized into intra-modality and inter-modality relationships. Intra-modality edges connect nodes within the same modality, capturing internal relationships specific to that modality. Inter-modality edges connect nodes across different modalities but at the same level of representation, enabling cross-modal interaction. The edge sets E are defined as follows:

$$E_{intra} = \bigcup_{i=1}^{n} \bigcup_{j=1}^{m} \bigcup_{k=1}^{m} (h_{ij}, h_{ik}) \tag{5.1}$$

$$E_{inter} = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} \bigcup_{k=1}^{n} (h_{ji}, h_{ki}) \tag{5.2}$$

$$E = E_{intra} \cup E_{inter} \tag{5.3}$$

This graph construction approach captures both modality-specific (intra-modality) relationships and cross-modal (inter-modality) dependencies. By linking all nodes within a modality, the intra-modality structure captures internal correlations and hierarchical dependencies inherent to that modality. On the other hand, the inter-modality connections allow the model to integrate information from different modalities at the same representational level, promoting the learning of complementary features.

Graph Learning

Once the graph $\mathcal{G}$ is constructed, we trained a GCN to jointly learn intra-modality and inter-modality relationships. The graph convolution operator as in [3]:

$$\widetilde{H}^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)})$$

where $\widetilde{A} = A + I_n$ denotes the adjacency matrix of the undirected graph $\mathcal{G}$ with inserted self-connections, I_n represents the identity matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ is the diagonal degree

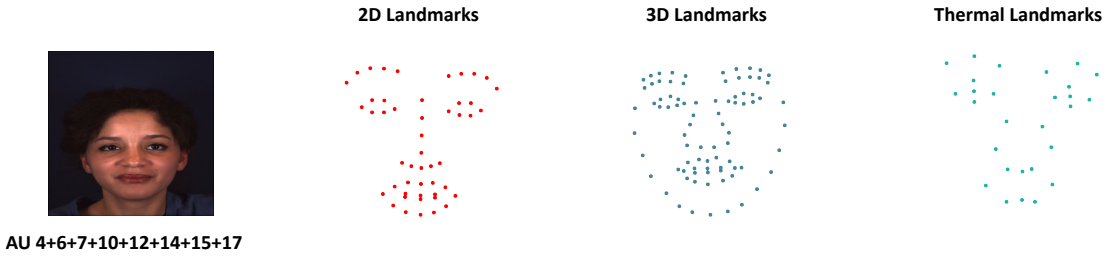


Figure 5.2: Plot of a subject's face from the BP4D+ dataset, including associated facial action units, 2D landmarks, 3D landmarks, and Thermal landmarks.

matrix, $W^{(l)}$ is a learnable weight matrix, and $\sigma(.)$ an activation function. H^l represents the matrix of activations in the l^{th} layer; $H^0 = X$, where X is the matrix of input node feature.

This graph-based learning mechanism allows our model to effectively capture and fuse information across modalities, leading to a richer and more nuanced feature space that is critical for addressing the heterogeneity and complexity inherent in multimodal data.

Pain Classifier

As previously discussed, the graph nodes are initialized with the feature representations extracted from each modality, denoted as $h^0_{11}, \ldots, h^0_{nm}$. After passing through k layers of the GCN, the encoded features $h^k_{11}, \ldots, h^k_{nm}$ are gathered into a single feature vector h^k:

$$h^k = [h^k_{11}, \ldots, h^k_{1m}, \ldots, h^k_{n1}, \ldots, h^k_{nm}]$$

These features are then passed through a global average pooling layer, followed by a fully connected neural network to predict the class label:

$$\hat{y} = fc(AvgPool(h^k))$$

Here, *AvgPool* denotes the global average pooling layer, *fc* represents the fully connected layer, and $\hat{y}$ is the predicted pain probability.

5.4 Experimental Results

5.4.1 Datasets

We conducted our experiments using the BP4D+ multimodal dataset [119], which comprises data from 140 subjects (58 male and 82 female) aged between 18 and 66 years old.

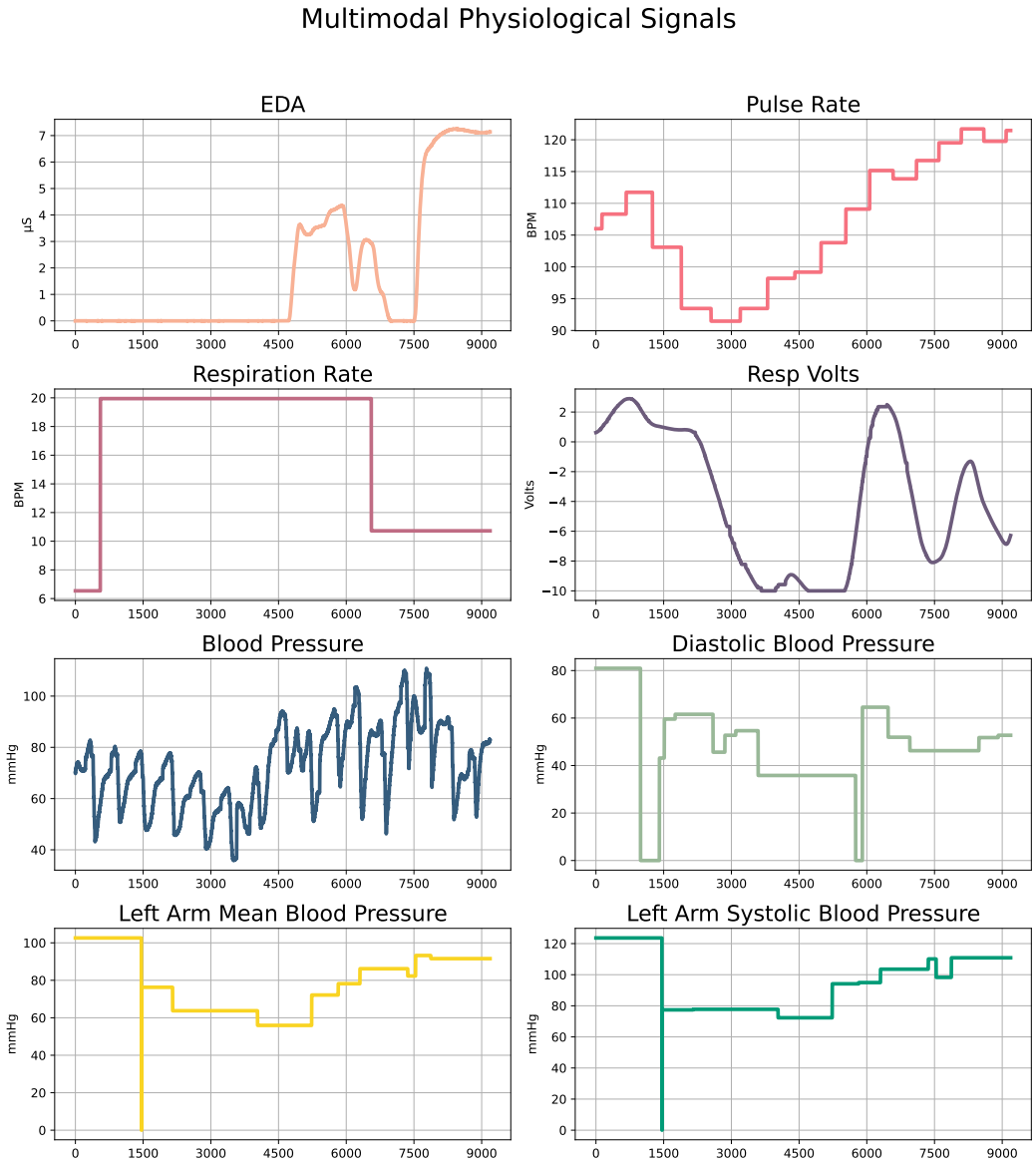


Figure 5.3: Plot of the electrodermal activity (EDA), pulse rate, respiration rate, respiration-related voltage measurement, blood pressure, diastolic blood pressure, left arm mean blood pressure, and left arm systolic blood pressure data for a subject from the BP4D+ dataset. Data is collected from the start to the end of the recording session, with the x-axis representing time steps. BPM stands for beats per minute for pulse rate and breaths per minute for respiration rate.

These subjects performed 10 tasks designed to elicit 10 authentic emotions. For detailed information about the specific tasks performed, please refer to Table 5.1. The dataset provides various modalities captured for each emotion elicitation task, including 3D face meshes, 2D RGB videos, thermal videos, facial landmarks (2D, 3D, and Thermal), and eight physiological signals: electrodermal activity, diastolic blood pressure, mean blood pressure, systolic blood pressure, raw blood pressure, pulse rate, respiration rate, and respiration voltage.

Table 5.1: Overview of the ten emotion elicitation tasks used in the BP4D+ dataset. The table details the activities performed by the participants and the corresponding emotions they are intended to evoke.

Task	Activity performed	Emotion elicited
T1	Listen to a humorous joke during an interview	Happiness
T2	Observe a 3D avatar of oneself	Surprise
T3	Video clip: 911 emergency phone call	Sadness
T4	Experience a sudden burst of loud sound	Surprise
T5	Respond to a true or false question in an interview	Skeptical
T6	Perform an improvised silly song	Embarrassment
T7	Face a simulated physical threat in a game	Fear
T8	Immerse hand in ice-cold water	Physical pain
T9	Complain about poor performance in an interview	Angry
T10	Encounter an unpleasant odor	Disgust

Furthermore, facial action units (AUs) were annotated for both occurrence and intensity by FACS experts for four specific emotion-elicitation tasks: happiness, embarrassment, fear, and pain. The annotations focused exclusively on the most facially expressive segments of the video recordings. In this study, we used only the data associated to the annotated most expressive frames associated with these four emotions.

Since this chapter centers on the automatic pain detection task, sequences eliciting pain were treated as the positive class, while sequences eliciting the other three emotions served as the negative class.

5.4.2 Data Preprocessing

For all the facial landmark data (2D, 3D, and Thermal), we calculated the Euclidean distances between all pairs of landmarks for each time step across the video sequence. This transformation has been shown to improve performance compared to using the raw landmark coordinates directly.

Regarding the physiological data, which was originally sampled at a frequency of 1000 Hz, we downsampled the signals to match the video frame rate of 25 frames per second (fps). Next, for all data modalities, we segmented the data into non-overlapping sliding windows, each containing 350 time steps (approximately 14 seconds). If a sequence contained fewer than 350 time steps, it was padded with the last available data point to reach the required length.

Table 5.2: Unimodal pain detection: comparison with a state-of-the-art method on the BP4D+ dataset.

	2D		3D		Thermal		AUs		Physio	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Wu et al. [105]	91.59	89.46	91.27	89.30	83.53	83.37	-	-	**83.24**	**82.42**
Transformer	**92.99**	**92.93**	**92.45**	**92.15**	**86.81**	**85.41**	**92.11**	**92.15**	81.81	80.17

5.4.3 Implementation Details

To determine the optimal hyperparameters of our model, we employ a grid-search strategy, considering the following hyperparameters and their respective values:

- Dimension of the linear projection layer: 256 and 512
- Number of multi-head attention: 4 and 8
- Number of Transformer encoder layers: 1, 2, 3, and 4

Regarding the GCN component in our framework, we employed two convolutional layers. The first layer had a hidden dimension of 512, and the second layer had a hidden dimension of 128. Additionaly, the batch size was fixed to 16, the maximum number of epochs to 500, and the learning rate to 10^{-5}. All models were trained using the Adam optimizer [113], with exponential decay rates for the first and second-moment estimates set at 0.9 and 0.999, respectively. The entire framework was implemented using PyTorch [114] and PyTorch Geometric [151].

5.4.4 Results

We conducted both unimodal and multimodal pain detection experiments using the BP4D+ dataset. In line with previous studies [148, 105], we employed a subject-independent 10-fold cross-validation strategy for model evaluation and calculated accuracy and the weighted average F1-score as evaluation metrics.

For clarity, we use the terms 2D, 3D, Thermal, and Physio to refer to 2D landmarks, 3D landmarks, Thermal landmarks, and physiological data, respectively. In multimodal cases, combination such as 2D + Physio represent the fusion of 2D landmarks and physiological data. This notation is applied consistently across other modality combinations.

Unimodal Pain Detection

In Table 5.2, we compare the performance of a Transformer encoder model with the method proposed by Wu et al. [105] for the task of pain detection using the following input modal-

Table 5.3: Multimodal pain detection using combination of two modalities on the BP4D+ dataset.

	2D + Physio		3D + Physio		Thermal + Physio		AUs + Physio	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Wu et al. [105]	93.45	91.37	92.66	90.47	89.07	88.96	-	-
MMT-early	92.65	92.63	90.99	90.74	85.67	84.49	90.79	90.43
MMT-inter	90.82	90.77	88.51	88.27	79.14	80.15	**94.06**	94.01
MMT-late	91.02	91.04	87.39	87.79	77.69	78.92	93.89	93.97
MMT-all	93.53	93.38	91.55	91.19	86.35	85.78	93.70	93.74
MMGT-intra	93.53	93.39	91.36	91.11	87.46	86.94	93.17	93.18
MMGT-light	93.01	92.94	92.44	92.29	87.64	87.15	93.36	93.32
MMGT (ours)	**93.90**	**93.82**	**93.72**	**93.59**	**89.45**	**89.14**	**94.07**	**94.10**

ities: physiological data, 2D, 3D, and Thermal facial landmarks.

The Transformer model surpasses Wu et al. method in terms of both accuracy and F1-score across the 2D, 3D, and Thermal facial landmarks modalities. Specifically, for 2D landmarks, the Transformer model achieved improvements of 1.40% in accuracy and 3.47% in F1-score. For 3D landmarks, the enhancements are 1.18% in accuracy and 2.05% in F1-score. For Thermal landmarks, the Transformer model achieves increases of of 3.28% in accuracy and 2.04% in F1-score. However, when using physiological data, the method by Wu et al. outperforms the Transformer model, with improvements of 1.43% in accuracy and 2.17% in F1-score.

Additionally, we reported the performance of the Transformer when employing the AUs modality. Among the different modalities tested, the best results were achieved using 2D landmarks, while physiological data yields the lowest results. These findings highlight the superiority of facial expression-based modalities for the task of pain detection in the BP4D+ dataset.

In the following, we present the results of our proposed MMGT architecture when integrating multiple modalities (combinations of two and three modalities) from those used in unimodal pain detection.

Multimodal Pain Detection

Our experiments with multimodal data are summarized in Tables 5.3, 5.4 and 5.5. In these tables, we evaluate our proposed MMGT model against machine learning [148], deep learning [152, 153], and geometric-based [131, 105] state-of-the-art methods. The pain detection approaches from [152, 153, 131] were reimplemented and reported in Table 5.3 by [105].

Table 5.4: Multimodal pain detection using combination of three modalities on the BP4D+ dataset.

	2D + AUs + Physio		3D + AUs + Physio		Thermal + AUs + Physio	
Method	Acc	F1	Acc	F1	Acc	F1
MMT-early	93.31	93.33	92.03	92.01	92.42	92.23
MMT-inter	93.17	93.31	93.62	93.59	93.66	93.67
MMT-late	93.64	93.79	92.76	92.92	93.62	93.74
MMT-all	94.09	94.00	93.66	93.53	93.39	93.43
MMGT-intra	93.52	93.52	93.51	93.36	92.86	92.85
MMGT-light	94.59	94.56	93.66	93.61	93.04	93.08
MMGT (ours)	**94.95**	**94.91**	**94.41**	**94.31**	**93.87**	**93.93**

We trained our MMGT architecture using different combinations of the modalities listed in Table 5.2. Initially, we combined physiological data with each of the other four facial expression-based modalities (AUs, 2D, 3D, Thermal landmarks), as shown in Table 5.3. For combinations involving three modalities, we combined AUs and physiological data with all types of facial landmarks, as detailed in Table 5.4.

Using two modalities, MMGT achieves state-of-the-art results across all tested combinations, as demonstrated in Tables 5.3 and 5.5. Notably, compared to Wu et al. [105], the most significant improvements are seen with the combination of 3D landmarks and physiological data. Our MMGT framework outperformed their approach by 1.06% in accuracy and 3.12% in F1-score, as shown in Table 5.3.

Moreover, it is important to highlight that the fusion of physiological data with other modalities consistently outperforms both unimodal approaches across all combinations, underscoring the synergistic benefits of integrating physiological signals with facial expression-based features for more accurate pain detection. Among all tested combinations of two modalities, our best results were obtained using AUs and physiological data, achieving 94.07% of accuracy and 94.10% of F1-score. Compared to Hinduja et al. [148], the only state-of-the-art method that also combines AUs and physiological data, we observe an improvement of 4.87% and 13.33% in terms of accuracy and F1-score relative to the pain class, respectively. For a fair comparison with [148], we reported the F1-score only for the pain class in Table 5.5. In contrast, to provide a fair comparison with other models, the remaining F1-scores reported in Tables 5.3, 5.4 and 5.5 are the weighted average F1-scores for both pain and non-pain classes.

Although the fusion of thermal landmarks and physiological data yields the lowest performance among the two-modality combinations, it still surpasses unimodal Transformer models trained solely on these modalities, demonstrating the added value of even less effective modality pairings.

Table 5.5: Comparison of our pain detection method with state-of-the-art results for the fusion of AUs and Physio, as well as the fusion of 2D and Physio. 'Early' and 'late' refer to the fusion strategies employed in these state-of-the-art methods, indicating whether the fusion occurs at an early or late stage in their frameworks.

Method	Acc	F1
AUs + Physio		
Hinduja et al. [148]	89.20	75.00
MMGT (ours)	**94.07**	**88.33**
2D + Physio		
Szczapa et al. (late) [131]	82.77	76.32
Szczapa et al. (early) [131]	84.32	78.83
Huang et al. (early) [152]	87.94	87.16
Huang et al. (late) [152]	89.36	89.13
Choo et al. (late) [153]	89.08	88.68
Choo et al. (early) [153]	89.80	89.46
Wu et al. [105]	93.45	91.37
MMGT (ours)	**93.90**	**93.82**

Given that our best combination of two modalities was achieved with AUs and physiological data, we tested whether we could further improve the MMGT performances by considering a third modality. This led to the combination of AUs, physiological data, and each type of facial landmark data (2D, 3D, and Thermal). As shown in Table 5.4, our hypothesis was validated in the cases where 2D and 3D landmarks were considered as a third modality, but not when thermal landmarks were included. The lower performance with thermal landmarks can be attributed to their weaker performance in the unimodal benchmarks, as seen in Table 5.2, where thermal landmarks yielded the poorest results among all facial landmark types for pain detection. This likely indicates that thermal landmarks provide fewer and less informative features for effective multimodal fusion.

Notably, the best overall performance was achieved with the combination of AUs, physiological data, and 2D landmarks using the MMGT architecture.

Ablation Study

As part of our ablation study, we found that integrating intermediate representations from unimodal Transformer encoders across multiple modalities using a GCN outperformed traditional multimodal fusion techniques, including early, intermediate, and late fusion ap-

proaches. In Tables 5.3 and 5.4, we present the results obtained with the aforementioned fusion techniques for different combinations of two and three modalities. Specifically, the Multimodal Transformer (MMT) approaches are defined as follows: MMT-early performs fusion at the input level through concatenation; MMT-inter concatenates the final representation layers from each Transformer and applies a fully connected layer for pain detection; and MMT-late averages the final decisions from each unimodal Transformer encoder.

As we can see in Tables 5.3 and 5.4, MMGT outperformed all fusion techniques for all combination of two and three modalities. We attribute this superior performance to two key factors: (1) the incorporation of intermediate representations from the Transformer encoders, and (2) the application of a GCN, which effectively fuses these representations to capture complex cross-modal patterns. Tables 5.3 and 5.4 show that, in general, traditional fusion techniques result in inferior performance compared to the best-performing modality used individually, which does not hold true for MMGT. For instance, fusing physiological data with 2D landmarks using traditional fusion techniques leads to a drop of at least 0.30% in terms of F1-score compared to 2D landmarks used alone (see Table 5.2). On the other hand, MMGT improves upon all individual modalities.

The superior performance of MMGT compared to traditional fusion techniques can be attributed to its ability to capture complex cross-modal interactions by integrating intermediate representations from each modality using a GCN. Unlike traditional methods, which often suffer from information loss or insufficient modeling of modality dependencies, MMGT effectively retains and leverages richer features across modalities. This approach leads to better generalization and avoids the performance drop commonly seen when fusing less complementary modalities.

To further examine the role of GCNs in efficiently combining intermediate representations from multiple modalities, we conducted an ablation study comparing MMGT with a Multimodal Transformer variant (MMT-all) that does not utilize graphs. Specifically, MMT-all directly concatenates the intermediate representations $h^0_{11}, \ldots, h^0_{nm}$, which are then fed into a fully connected layer for pain detection. As shown in Tables 5.3 and 5.4, MMGT consistently outperforms MMT-all for nearly all two- and three-modality combinations. For example, when using 3D landmarks and physiological data, we observed improvements of 2.17% in accuracy and 2.40% in F1-score, respectively.

Graph Construction Variations

Furthermore, we explore the impact of different graph construction strategies within our MMGT framework by evaluating two specific variants: MMGT-intra and MMGT-light. Each variant adopts a distinct approach to constructing the graph that models the connections among the extracted feature representations from the different modalities.

The MMGT-intra variant focuses exclusively on intra-modality interactions by restricting edges to connections within a single modality. By isolating modality-specific relationships, MMGT-intra aims to capture detailed interactions within each modality. This approach contrasts with the original MMGT architecture, which also includes connections between nodes from different modalities at similar representation levels. On the other hand,

the second variant, MMGT-light, simplifies the overall graph structure by connecting only the final representations of each modality. This variant links nodes across modalities but does so only at the highest level of abstraction.

Both MMGT-intra and MMGT-light demonstrate competitive performance, surpassing traditional fusion techniques for most modality combinations. For instance, MMGT-intra achieved an accuracy of 93.53% and an F1-score of 93.39% when combining 2D landmarks and physiological data, closely matching the original MMGT's performance (93.90% accuracy and 93.82% F1-score). Similarly, MMGT-light performed robustly across different modality combinations, showing that even a simplified graph structure can retain most of the benefits of multimodal integration.

However, despite these strengths, MMGT consistently outperforms both variants across all two- and three-modality combination settings, demonstrating the importance of both intra- and inter-modality interactions in capturing the full spectrum of multimodal relationships for optimal pain detection performance. For instance, when combining AUs and physiological data, MMGT achieves an accuracy of 94.07% and an F1-score of 94.10%, surpassing MMGT-intra's 93.52% accuracy and 93.45% F1-score, as well as MMGT-light's 93.53% accuracy and 93.41% F1-score.

Furthermore, the superior performance of MMGT compared to MMGT-light highlights the value of leveraging intermediate feature representations through our proposed multimodal fusion framework. MMGT-light also outperforms all classical fusion techniques for various combinations, including each type of facial landmark with physiological data, as well as 2D + AUs + Physio and 3D + AUs + Physio combinations. This further highlights the effectiveness of the multimodal GCN module.

Our experiments reveal several key insights for graph-based multimodal fusion in pain detection. First, the performance gap between MMGT and its variants underscores the importance of capturing both intra- and inter-modality dependencies. Additionally, the MMGT architecture, which leverages intra- and inter-modality connections, consistently delivers state-of-the-art results, demonstrating that a balanced approach to graph construction leads to superior performance. Although MMGT-intra and MMGT-light provide useful insights into simplifying graph design, the comprehensive integration strategy employed by MMGT remains the most effective for achieving the best performances.

In summary, the overall findings reveal three key insights for multimodal learning with our MMGT framework for the task of pain detection: (1) Integrating different modalities (e.g., combining physiological data with facial expressions) consistently improves pain detection performance, (2) Adding more modalities generally leads to enhanced classification results, and (3) The way multimodal interactions are handled within the graph structure significantly influences the final task performance.

5.5 Discussion

The results from this chapter highlight the potential of the MMGT framework in advancing the field of automatic pain detection and, more broadly, emotion recognition. This discussion section delves into the implications of these findings, the limitations of our work, and potential directions for future research.

5.5.1 Implications of Findings

The superior performance of our proposed MMGT architecture underscores the importance of considering the multimodal nature of pain and designing specialized architectures to effectively leverage the strengths of each modality. By integrating physiological signals and facial expression data, our MMGT captures a more comprehensive representation of the pain experience.

Furthermore, the MMGT's enhanced detection accuracy highlights the effectiveness of graph-based models in capturing complex relationships among modality-specific intermediate feature representations. This increased accuracy could be particularly valuable in clinical settings, where precise pain assessment can directly influence patient care.

While our primary focus was pain detection, our study also contributes to the broader field of emotion recognition, especially given that the BP4D+ dataset encompasses multiple emotional states. The versatility of our MMGT model extends beyond pain detection, as it effectively distinguishes between pain and other emotional states. This capability for emotion recognition holds promising applications across various domains, including healthcare, human-computer interaction, education, among others.

One particularly important application lies in medical simulations training, where recognizing a range of emotional states could enhance training programs by providing real-time feedback on medical students' emotional states, allowing for the dynamic adjustment of training scenarios. The emotion recognition system will process multimodal data, including facial expressions and physiological signals from students during simulations, to accurately predict their emotional states.

5.5.2 Limitations

Despite the promising results, several limitations need to be acknowledged.

The experiments were conducted on the BP4D+ dataset, which, while comprehensive in terms of data modalities, may not fully represent the diversity of pain and emotion expressions across different populations. This dataset predominantly features controlled environments with a specific demographic composition, potentially limiting the generalizability of our findings. Therefore, cross-dataset validation is necessary to evaluate the generalizability of the MMGT across different populations and data collection protocols. More diverse datasets representing a broader range of individuals and real-world conditions would pro-

vide a better foundation for training models applicable to more varied contexts.

Labeling of pain and other emotional states in BP4D+ is elicited through specific tasks designed to induce controlled emotional expressions, which may not fully reflect the complexity of naturalistic pain and emotional states encountered in real-world settings. While these controlled settings offer consistency in model training, the induced emotions may not capture the subtle variations and spontaneous expressions seen in everyday situations.

While our MMGT model demonstrates high accuracy within the controlled environment of the BP4D+ dataset, transitioning to real-world scenarios remains challenging. Differences in lighting conditions, camera angles, facial occlusions, and individual pain expression variations can significantly affect the model's performance. These uncontrolled variables in real-world applications, require the model to be robust enough to handle visual noise and incomplete data.

The use of multiple modalities, such as physiological signals and facial expressions, can improve the richness of data and enhances pain detection and emotion recognition accuracy. However, it introduces several challenges, particularly during inference. One of the primary issues is the availability and synchronization of multimodal data. In real-world applications, it may not always be possible to capture all modalities simultaneously. During inference, the absence of certain modalities due to equipment failure, or environmental limitations could render the model inoperable.

The way we constructed the graph in the MMGT model introduces certain limitations. While the graph structure allows for capturing relationships between different modality-specific features, the current design may not fully optimize the interactions between modalities, particularly when the connections between nodes are predefined rather than learned dynamically. Additionally, our framework assumes that for each modality we extracted the same number of intermediate representations. However, this approach may not be optimal, as different modalities might achieve the best performance with varying numbers of Transformer layers.

The MMGT, like deep learning models in general, is considered a black box model due to its complexity and the non-transparent nature of their prediction processes. Although the combination of Transformer and GCN offers high pain detection accuracy, the complex architecture of these models makes it challenging to understand how they arrive at specific predictions. Clinicians and other end-users may be hesitant to adopt a model they cannot fully interpret, particularly in sensitive applications such as pain detection, where decisions have a direct impact on patient care.

Lastly, our MMGT framework, which incorporates Transformer encoders and GCN, is computationally intensive. This complexity may limit its real-time applicability in clinical settings or in other type of environment who would benefit from pain and emotion recognition, without significant computational resources.

5.5.3 Future Directions

The promising results of this study open several avenues for future research.

Expanding the dataset to include a wider range of pain expressions from diverse demographic groups will improve the model's robustness and generalizability. Cross-dataset validation is crucial for evaluating the model's performance in real-world settings. By training and testing the MMGT on multiple datasets with varied conditions and populations, we can better assess its adaptability and consistency across diverse environments.

To address the challenge of controlled versus naturalistic pain expressions, future research should focus on integrating datasets that capture more spontaneous and nuanced emotional expressions. Incorporating data from real-world scenarios and less controlled environments can provide a more comprehensive understanding of pain and emotion, helping to bridge the gap between controlled experimental conditions and real-world applications.

Improving the graph structure of the MMGT model is another key area for future work. Dynamic graph construction, where the model learns optimal connections between modality-specific features, could enhance the fusion of intra- and inter-modality information. Incorporating attention mechanisms within the graph could enable the model to focus on the most relevant interactions, potentially leading to more accurate pain and emotion detection.

Personalization techniques should also be considered to account for individual differences in both pain and emotional expression. Models could learn from individual person over time to better tune to personal pain and emotion indicators, thereby improving accuracy. Incorporating adaptive learning mechanisms would allow the model to continuously refine its predictions based on feedback and evolving patterns in individual behavior.

In future research, enhancing model interpretability will be essential to increasing trust and usability in real-world applications. Developing techniques to make the MMGT model's decision-making process more transparent could facilitate its adoption by clinicians and end-users. For instance, integrating explainable AI methods could help elucidate how the model arrive at its predictions. Visualization tools that highlight which features or modalities most influence the model's outputs could provide valuable insights.

Lastly, addressing the computational intensity of the MMGT framework is crucial for practical deployment. Research could focus on optimizing the model's computational efficiency, such as developing lightweight versions of the model to enable real-time applications in clinical settings and other environments that would benefit from pain and emotion recognition.

5.6 Conclusion

In this chapter, we introduced the Multimodal Graph-based Transformer (MMGT), a novel multimodal fusion framework designed for the task of pain detection. The MMGT effectively leverages the strengths of both Transformer encoders and GCN to integrate multiple data modalities, capturing the complex relationships within and across modalities. We performed extensive benchmark on the BP4D+ dataset, demonstrating state-of-the-art performance across various modality combinations, particularly when integrating 2D facial

landmarks, facial action units, and physiological data.

Our results highlight several important insights. First, the MMGT's ability to consistently outperform single-modality models and even combinations of two input modalities underscores the complementary nature of the data sources employed. This reflects the inherent complexity of pain as a multimodal experience, where physiological signals and facial expressions together provide a more comprehensive picture than any one modality alone. The synergy between these modalities, captured effectively by the MMGT, illustrates the potential of sophisticated fusion techniques in improving automatic pain detection.

However, despite these promising results, several challenges remain. The model's generalizability across diverse datasets and its adaptability to varied, real-world environments are still open questions that require further investigation. Additionally, the computational demands posed by the integration of Transformer and GCN architectures may limit the MMGT's practical deployment in resource-constrained settings. Addressing these limitations points toward important directions for future research, such as refining the model to operate more efficiently and exploring adaptive, personalized approaches that can account for individual variations in pain perception and emotional expression.

In summary, this work demonstrates that by leveraging advanced multimodal fusion strategies, we can significantly enhance pain detection models, bringing us closer to more reliable, context-aware systems. Beyond this specific application, the MMGT framework offers broader implications for advancing emotion recognition technologies, with potential benefits spanning healthcare, human-computer interaction, and beyond.

Part II

Surgical Data Science

Chapter 6

STGFormer: Spatial-Temporal Graph Transformer for Surgical Skill Assessment

Contents

This chapter presents technical contributions to the field of surgical skill assessment. We propose a novel deep learning framework that combines a Graph Convolutional Network (GCN) with a Transformer encoder for the task of surgical skill assessment. The GCN has been designed to learn spatio-temporal representations from hand skeleton sequences, while the Transformer encoder captures long-range dependencies within these representations. The objective of the proposed framework is to accurately differentiate sequences of movements of attending surgeons from those of surgical residents based on the analysis of their hand skeleton sequences, thereby identifying expert-level movements from novice actions. We tested our framework on two collected surgical simulation tasks: circular cutting and needle passing.

In Section 6.1, we discuss the task of surgical skill assessment, outline its challenges and highlight the need for automated solutions in this domain. Section 6.2 provides a detailed review of current state-of-the-art methods, focusing on approaches that employ kinematics and video data for surgical skill evaluation. In Section 6.3, we introduce our proposed framework. Section 6.4 presents the two surgical simulation datasets collected for evaluating our framework. Section 6.5 outlines the data preprocessing steps, the evaluation framework, and our experimental results. Lastly, in Section 6.6, we discuss the implications of our findings, the limitations of our work, and potential future research directions. Section 6.7 concludes the chapter by summarizing our contributions.

6.1 Introduction

Assessing surgical skill is a crucial aspect of surgical education and ongoing professional development. This process involves evaluating and measuring a surgeon's technical proficiency and competence in performing surgical procedures. Traditionally, evaluation has depended largely on subjective assessments by experienced surgeons, utilizing both global and task-specific checklists [154, 155]. However, these assessment methods present multiple limitations, such as being biased towards the evaluator, being time-consuming, and lacking standardization across various surgical tasks.

As surgical procedures become increasingly complex, there is a growing demand for objective methods to assess surgical skills, particularly through simulation-based training. Consequently, the development of tools for evaluating surgical skills during the performance of surgical tasks has gained significant attention.

Automated assessment systems present several key advantages. Firstly, automated systems can provide real-time feedback on performance, allowing practitioners to quickly identify errors and areas for improvement. Secondly, these systems can enable the tracking of a practitioner's learning progress over time, fostering continuous skill development. Thirdly, integrating automated assessment into simulations will ensure that trainees achieve the required competencies before transitioning to actual surgeries. Finally, the scalability of these systems will allow for the simultaneous assessment of multiple trainees, significantly enhancing the overall training process.

To develop these automated systems, a wide range of data sources can be employed, in-

cluding kinematic data, instrument trajectories, and video analysis. By analyzing these data modalities, automated assessment tools can deliver comprehensive feedback on various aspects of surgical performance, such as expertise level, speed, and dexterity. Expertise level, for instance, can be quantified numerically using metrics like the average OSATS score [154] or categorized into novice and expert levels, providing clear, objective measures of surgical competence.

In recent years, GCNs have become the de facto choice for modeling relational data due to their ability to capture both the local and global structure of graphs. This has led to GCNs achieving state-of-the-art performance across various tasks involving spatiotemporal data, where the data can be effectively represented as graphs. These tasks include traffic forecasting [156, 157], weather forecasting [158], action recognition [159], and gesture recognition [160]. Similarly, the Transformer architecture [2] has fundamentally transformed the landscape of natural language processing (NLP), setting new benchmarks across various NLP tasks. Beyond NLP, Transformers have also achieved state-of-the-art performance in a wide range of other domains, including skeleton-based action recognition, where they have demonstrated superior capabilities in capturing spatiotemporal dependencies and complex patterns [161, 162, 163].

Build on the success of GCNs and Transformers in processing spatiotemporal data, particularly skeleton-based data, this chapter explores the potential of using hand skeleton sequences for surgical skill assessment. We propose the STGFormer framework, which leverages the graph structure of hand skeletons and the dynamic spatiotemporal patterns inherent in hand movement sequences. By combining the strengths of GCNs for learning spatial-temporal representations of hand skeleton sequences with the Transformer encoder's ability to capture long-range dependencies, the STGFormer framework models effectively interactions between hand joints across time for assessing surgical proficiency. To the best of our knowledge, our approach represents the first attempt to evaluate surgical proficiency using hand skeleton sequences, presenting a novel and scalable approach to skill assessment.

To validate the use of hand skeleton sequences and our proposed STGFormer framework for surgical skill assessment, we collected two novel surgical simulation datasets, each featuring a different surgical task performed by both attending surgeons and surgical residents. The first dataset includes a circular cutting task, while the second involves a needle passing task. Both tasks were performed using the VirtaMed medical simulator.

The contributions of this work are three-fold and can be summarized as follows:

1. We propose the use of hand skeleton sequences for the task of surgical skill assessment.

2. We present two newly collected simulation datasets for surgical skill assessment, featuring performances by attending surgeons and surgical residents on two fundamental tasks.

3. We introduce a novel deep learning framework that captures the dynamic spatial-temporal correlations of hand skeleton sequences. Our framework demonstrates state-of-the-art performance on the two collected datasets.

6.2 Related Work

Since the most common data sources for assessing surgical skills are instrument motion analysis and video data, we will divide our literature review into two sections, each focusing on one of these data modalities.

6.2.1 Robotic Kinematics-Based Assessment

Surgical skill assessment has traditionally relied on instrument motion data as a key feature for evaluating the proficiency of surgeons. Instrument motion data, which includes the trajectories and movements of surgical tools, provides valuable insights into the the practitioner's skills when performing a surgical procedure.

In this regard, the JIGSAWS (JHU-ISI Gesture and Skill Assessment Working Set) [164] dataset has been introduced as a benchmark for evaluating surgical skill assessment methods. This dataset comprises data from eight attending surgeons with varying levels of experience, each performing five trials of three fundamental surgical procedures: suturing, knot tying, and needle passing, using the da Vinci Surgical System. It includes kinematic data, capturing the positions, and velocities of the robotic instruments for both the right and left hands, alongside video recordings of the surgical procedures.

Early approaches for the task of surgical skill assessment using instrument motion data often relied on handcrafted features and statistical models. One early study introduced the sparse Hidden Markov Model (sparse HMM) as a variant of the traditional Hidden Markov Model for skill evaluation [165]. More recently, Fard et al. [166] proposed computing eight global movement features: task completion time, path length, depth perception, speed, motion smoothness, and curvature, along with two additional features they introduced: turning angle and tortuosity. Subsequently, these features were employed to train three classification models: k-nearest neighbors, logistic regression, and support vector machines for binary classification to distinguish between expert and novice performance. In another study, Zia et al. [167] proposed computing a range of features including Sequential Motion Texture (SMT), Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Approximate Entropy (ApEn). Each feature set was individually processed using Principal Component Analysis (PCA) before being employed for surgical skill classification. The study also introduced a novel weighted feature fusion technique, which combined predictions derived from each feature set. This fusion technique involved solving a least squares equation.

Recent advancements in deep learning have greatly enhanced the modeling and analysis of instrument motion data, eliminating the need for manually engineered features. These techniques leverage end-to-end learning to extract skill-related information directly from the data. Several studies [168, 169, 170] have utilized Convolutional Neural Networks (CNNs) to classify surgical skills using raw kinematic data. Additionally, other research [171, 172] has explored hybrid architectures that combine CNN and Recurrent Neural Network (RNN) branches for spatial and temporal feature learning, respectively. For example, Wang et al. [171] proposed the SATR-DL framework, which features a dual-branch struc-

ture: one branch employs a CNN to extract spatial features, while the other utilizes a Gated Recurrent Unit (GRU) to capture temporal dynamics. The outputs of these branches are concatenated and employed for skill assessment.

While kinematic data is a valuable tool for assessing surgical skills, it comes with several limitations. A significant drawback is the reliance on specialized equipment, such as surgical robotic systems or advanced technologies, which are often costly and restrict access to well-resourced settings. This limits the feasibility of using kinematic assessments in less equipped environments. Additionally, kinematic data collection in robotic systems is confined to robot-assisted procedures, excluding many surgeries or surgical simulation procedures performed manually or with minimal robotic support. Moreover, kinematic data primarily measures the movements and trajectories of surgical instruments, providing limited contextual insight. It fails to capture critical elements of surgical performance, such as tissue handling, hand movements, and the surgeon's interaction with the surgical environment. This narrow scope can result in an incomplete assessment of a surgeon's overall skill. Furthermore, the accuracy of kinematic data can be affected by sensor inaccuracies and environmental noise, leading to potential errors in evaluating surgical proficiency.

6.2.2 Video-Based Assessment

In addition to instrument motion data, video recordings have emerged as a powerful source of information for surgical skill assessment. Videos capture the entire surgical scene, including instrument motion, tissue interactions, and overall procedural context, providing a holistic view of the surgery that kinematic data alone cannot achieve.

Several studies proposed to computed image features for video-based surgical skill assessment. For instance, in [173] and [174], the authors proposed computing spatiotemporal interest points (STIPs) to identify key regions in the images of a video of the surgical procedure, subsequently calculated relevant descriptors. Specifically, in [174], the authors computed three descriptors for each STIP: the histogram of oriented gradients (HoG), the histogram of optical flow (HoF), and motion boundary histograms (MBHs). Next, they clustered the STIP descriptors within a video using a k-means algorithm to obtain a visual feature vocabulary and applied TF-IDF to obtain the BoW features. Furthermore, to address the lack of temporal information in BoW, the authors proposed Augmented Bag of Words (Aug. BoW) features. Additionally, they computed several other features, including the Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Sequential Motion Texture (SMT), Approximate Entropy (ApEn), and Cross Approximate Entropy (XApEn). For classification between experts and novices based on the extracted features, they employed logistic regression, linear SVM, and multilayer perceptron (MLP) classifiers.

Recently, there has been a shift toward deep learning-based methods, which can directly learn complex features from raw video data and achieve state-of-the-art performance in assessing surgical skills. In Hira et al. [174], in addition to the previously computed image features, they proposed using a Temporal Convolutional Network (TCN) on predict instrument keypoints for surgical skill assessment. They also introduced a CNN-LSTM framework, employing ResNet [67] for image feature extraction and LSTM to learn tempo-

ral dynamics. Furthermore, they enhanced both the ResNet image encoder and the LSTM with spatial and temporal attention mechanisms, respectively. In another study, Funke et al. [175] proposed using a Temporal Segment Network [176], which involves fine-tuning a pre-trained 3D Convolutional Neural Network to classify sequences of frames from a video. The predictions from these sequences are then aggregated using a consensus operator to determine the overall video classification result. Liu et al. [177] introduced a unified multi-path framework for automatic video-based surgical skill assessment, which considers various aspects of surgical skills including surgical tool usage, intraoperative event patterns, and other skill proxies. To model the relationships between these factors, a path dependency module was specially designed.

Video-based assessment methods offer several advantages over kinematic data alone. They provide richer contextual information, including tissue handling techniques, decision-making processes, and overall surgical strategy. Moreover, video recordings are non-intrusive and can be obtained using standard surgical cameras, making them applicable across a wide range of surgical procedures and settings. However, challenges remain, such as variability in camera viewpoints, lighting conditions, and occlusions, which can affect the accuracy and robustness of automated video analysis methods.

Unlike existing approaches, we propose assessing surgical skills through hand skeleton sequences. To achieve this, we have collected two new surgical simulation datasets featuring both Experts (attending surgeons) and Novices (surgical residents) performing the following tasks: circular cutting and needle passing. Existing publicly available datasets, such as JIGSAWS, often feature a limited number of subjects, restricting the generalizability of current approaches. In contrast, our datasets include a larger number of both Experts and Novices, enhancing the robustness of our findings. Additionally, we introduce a novel deep learning framework, STGFormer, designed to leverage hand skeleton data sequences for the task of surgical skill assessment. STGFormer integrates a GCN with a Transformer encoder to effectively learn spatial-temporal patterns from these sequences, enabling the differentiation of movement sequences characteristic of Experts versus Novices during the execution of surgical tasks.

6.3 Proposed Approach

This section presents our STGFormer framework, which is illustred in Figure 6.1. Our framework consists of two essential components: (1) a Spatial-Temporal Graph Transformer, comprising a GCN responsible for learning spatial-temporal representations from hand skeleton sequences, alongside a Transformer encoder for capturing global temporal patterns among the previously extracted representations; and (2) a Surgical Skill Classifier, which classifies the representations generated by the Spatial-Temporal Graph Transformer into Expert or Novice categories.

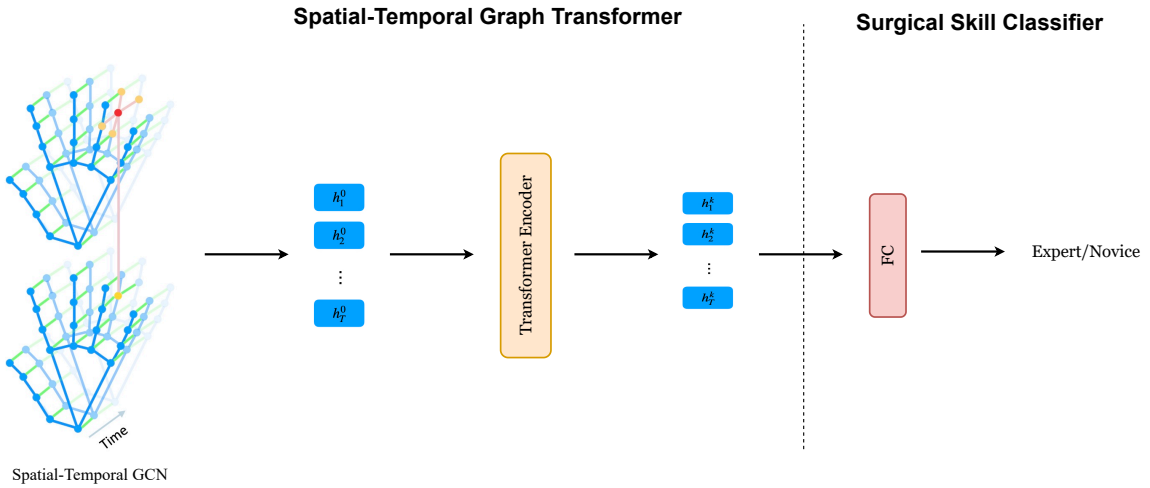


Figure 6.1: Illustration of our STGFormer-based surgical skill assessment framework, which is composed of two key components: Spatial-Temporal Graph Transformer and Surgical Skill Classifier.

6.3.1 Spatial-Temporal GCN

To extract higher-level feature representations from hand skeleton sequences, we first constructed spatial-temporal graphs based on the connectivity of hand joints. We propose two types of graph constructions: one using landmarks from a single hand (either left or right) and the other using landmarks from both hands. A GCN is then applied to process these graphs. These steps are detailed in the following.

Graph Construction

We constructed an undirected spatial-temporal graphs $\mathcal{G} = (V, E)$ to capture both spatial and temporal relationships between hand joints over a sequence of frames. Two types of graphs are developed: one for a single hand and an extended version for both hands.

One Hand: In this configuration, the graph $\mathcal{G} = (V, E)$ consists of N joints per hand over T frames. The set of nodes, V, represents the hand joints across all frames in the sequence, while the set of edges, E, represents the temporal and spatial connections between these joints across the different frames of the sequence. The construction process is as follows:

- **Nodes:** the nodes in the graph consist of all joints in the hand skeleton sequence, expressed as $V = \{v_{ti} \mid t = 1, \ldots, T, i = 1, \ldots, N\}$. Each node v_{ti} is initialized with its 3D coordinate information. The number of joint per hand, N, is equal to 21.

- **Edges:** the set of edges E is defined as the union of intra-skeleton connections, E_{intra}, and inter-frame connections, E_{inter}, in the graph, defined as follows:

$$E_{intra} = \{v_{ti}v_{tj} \mid (i,j) \in H, t \in \{1,\ldots,T\}\} \tag{6.1}$$

$$E_{inter} = \{v_{ti}v_{(t+1)i} \mid i \in \{1,\ldots,N\}, t \in \{1,\ldots,T-1\}\} \tag{6.2}$$

$$E = E_{intra} \cup E_{inter} \tag{6.3}$$

In Equation. 6.1, H represents the set of naturally connected hand joints.

Two Hands: The two-hand graph construction extends the single-hand setup by including inter-hand connections to capture interactions between the left and right hands. The process is outlined as follows:

- **Nodes:** the nodes in the graph consist of all joints in the hand skeleton sequence across the two hands, expressed as $V = \{v_{ti} \mid t = 1,\ldots,T, i = 1,\ldots,2N\}$. Each node v_{ti} is initialized with its 3D coordinate information. The number of joints per hand, N, is equal to 21, resulting in a total of $2N = 42$ nodes for both hands.
- **Edges:** the set of edges E is defined as the union of intra-skeleton connections, E_{intra}, inter-frame connections, E_{inter}, and inter-hand connections, E_{cross}, in the graph, defined as follows:

$$E_{intra} = \{v_{ti}v_{tj} \mid (i,j) \in H, t \in \{1,\ldots,T\}\} \tag{6.4}$$

$$E_{inter} = \{v_{ti}v_{(t+1)i} \mid i \in \{1,\ldots,2N\}, t \in \{1,\ldots,T-1\}\} \tag{6.5}$$

$$E_{cross} = \{v_{ti}v_{t(N+i)} \mid i \in \{1,\ldots,N\}, t \in \{1,\ldots,T\}\} \tag{6.6}$$

$$E = E_{intra} \cup E_{inter} \cup E_{cross} \tag{6.7}$$

Equation 6.6 introduces inter-hand edges that connect each joint in the left hand with its corresponding joint in the right hand at the same frame t, enabling the model to capture interactions between both hands.

Graph Learning

We trained spectral deep GCNs using the previously constructed graph $\mathcal{G}$. The graph convolution operator is defined as described in [3]:

$$\widetilde{H}^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{6.8}$$

Here, $\widetilde{A} = A + I_n$ represents the adjacency matrix of the undirected graph $\mathcal{G}$ with added self-connections, where I_n is the identity matrix. The diagonal degree matrix $\widetilde{D}$ is defined as $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$. The matrix $W^{(l)}$ is a layer-specific learnable weight matrix, and $\sigma(\cdot)$ denotes an activation function. The matrix $H^{(l)}$ represents the activations at the l^{th} layer, with the initial activations H^0 corresponding to the input node feature matrix X.

6.3.2 Transformer Encoder

To capture complex temporal dependencies in hand skeleton sequences, we input the high-level representations obtained from the GCN into a Transformer encoder. For each frame, joint representations j_{ti} are concatenated into a vector h_t^0 (Equation 6.9), and these vectors are aggregated into h^0 (Equation 6.10), which serves as the input for the Transformer encoder.

$$h_t^0 = [j_{t1}, j_{t2}, \ldots, j_{tN}] \tag{6.9}$$

$$h^0 = [h_1^0, h_2^0, \ldots, h_T^0] \tag{6.10}$$

After the forward pass through the k-th Transformer encoder layer, the model output a new representation h_i^k for each timestep. All these new representations are combined into a vector h^k.

$$h^k = [h_1^k, h_2^k, \ldots, h_T^k] \tag{6.11}$$

6.3.3 Surgical Skill Classifier

The final representation h^k (Equation 6.11), is fed into a fully connected neural network for surgical skill classification, predicting whether the hand skeleton sequence corresponds to an Expert or Novice skill level.

6.4 Surgical Simulation Datasets

This section introduces two surgical simulation datasets designed for benchmarking surgical skill assessment methods using hand skeleton sequences. The datasets were collected at the PRESAGE (Plateforme de Recherche et d'Enseignement par la Simulation pour l'Apprentissage des Attitudes et des Gestes) medical simulation center, affiliated with the Faculty of Medicine at the University of Lille.

The datasets consist of two distinct surgical simulation tasks: circular cutting and needle passing. These tasks were performed by both attending surgeons and surgical residents

Figure 6.2: The VirtaMed simulator.

using the VirtaMed medical simulator. The VirtaMed simulator is a state-of-the-art virtual reality training tool that provides high-fidelity graphics and haptic feedback, offering a highly realistic and immersive environment for medical professionals to practice various surgical procedures without involving live patients. Figure 6.2 depicts the VirtaMed simulator.

For both datasets, the protocol was approved by the Institutional Review Board of the University of Lille under reference number 2022-626-S108.

6.4.1 Circular Cutting Dataset

We collected side-view video recordings of hand movements from 16 participants, including 4 attending surgeons and 12 surgical residents from diverse surgical specialties, while they performed a circular cutting task on the VirtaMed simulator. Subsequently, the hand skeletons from the right hand were extracted from the videos using the method described in [178]. We focused exclusively on right-hand skeletons due to the unreliable detection of left-hand skeletons and its minor role in the circular cutting task. Figure 6.3 shows a participant performing the circular cutting task, alongside the right skeleton detected. Figure 6.4 illustrates the circular cutting task on the simulator screen, and Figure 6.5 presents the surgical tools associated with the simulator that were used to complete the task.

Figure 6.3: A side view of a participant performing the circular cutting task on the VirtaMed simulator.

The task is divided into multiple steps. First, participants entered the virtual environment using the laparoscope provided by the VirtaMed simulator and adjusted the view as needed. The laparoscope used is shown in Figure 6.5a. Next, participants utilized the atraumatic grasper tool, depicted in Figure 6.5b, to apply tension, and then used scissors, also shown in Figure 6.5b, to cut between the two circles.

6.4.2 Needle Passing Dataset

We recorded frontal-view videos of hand movements from 7 attending surgeons and 22 surgical residents from various surgical specialties while they performed a needle passing surgical task using the VirtaMed simulator. Hand skeletons for both the right and left hands were extracted using the method described in [178]. Figure 6.6 shows a subject performing the needle passing task, along with the detected hand skeletons of the right and left hands.

The needle passing task comprises several steps. Firstly, participants enter the virtual environment with the laparoscope and adjust the view to correctly position for the main phase of the exercise. Next, participants use their dominant hand to grasp the needle at the designated mark and pass it through the active ring without touching its edges. They then use their non-dominant hand to grasp the tip of the needle at the marked point and pull it through the ring. This sequence is repeated with additional rings. An illustration of the exercise is shown in Figure 6.7.

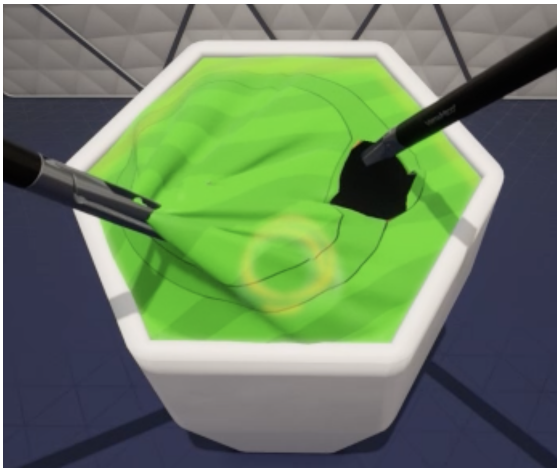


Figure 6.4: Illustration of the circular cutting task, captured from the VirtaMed simulator screen.

6.5 Experimental Results

6.5.1 Data Preprocessing

Hand Landmarks Normalization

To standardize the hand skeleton data and reduce variability due to initial hand positioning, we normalized each hand skeleton sequence. The normalization was performed by subtracting the coordinate of the first wrist joint v_{00} from the coordinates of all other joints in the hand skeleton sequence. Mathematically, this is represented as:

$$v'_{ij} = v_{ij} - v_{00}$$

where v_{ij} denotes the spatial coordinate of the i-th joint at the j-th time step, and v_{00} is the coordinate of the wrist joint at the 0-th time step. This normalization anchors the hand movements relative to the wrist, providing a consistent reference point and reducing the impact of different starting positions.

Segmentation

After the normalization step, we subdivide each sequence of hand skeleton data into non-overlapping sliding windows of 20 seconds, which was found optimal for both datasets. With a video frequency rate of 30 frames per second, this result in approximately 600 timesteps for each sub-sequence. As a result, we have a varying number of data sequences for each participant, which are directly dependent on the time taken to complete the surgical task.

While segmenting the sequences into windows introduces variability between them,

(a)

(b)

Figure 6.5: (a) Laparoscope; (b) Atraumatic Grasper / Scissors.

this approach offers several advantages. It makes the data more manageable and ensures consistent input sizes for our proposed framework. Additionally, it effectively increases the size of both datasets, which is particularly beneficial given the limited number of original sequences. Despite the inherent variability between windows, our proposed STGFormer framework achieves satisfactory results, as we will demonstrate later. This indicates that the model generalizes well across different types of movements involved in completing either surgical task.

6.5.2 Implementation Details

To determine the optimal hyperparameters for our STGFormer framework, we employ a grid-search strategy, exploring the following hyperparameters and their respective values:

- Dimension of the linear projection layer: 256 and 512
- Number of multi-head attention: 4 and 8
- Number of Transformer encoder layers: 1 and 2
- Number of convolution layer in the GCN: 1 and 2
- Dimension in each convolution layer: 8, 16, 32, and 64

Additionally, we varied the batch size between 16, 32, and 64. The maximum number of epochs was set to 150, with an early stopping criterion of 30 epochs. The learning rate was fixed at 10^{-4}. All models were trained using the Adam optimizer [113], with exponential decay rates for the first and second moment estimates set at 0.9 and 0.999, respectively. The entire framework was implemented using PyTorch [114].

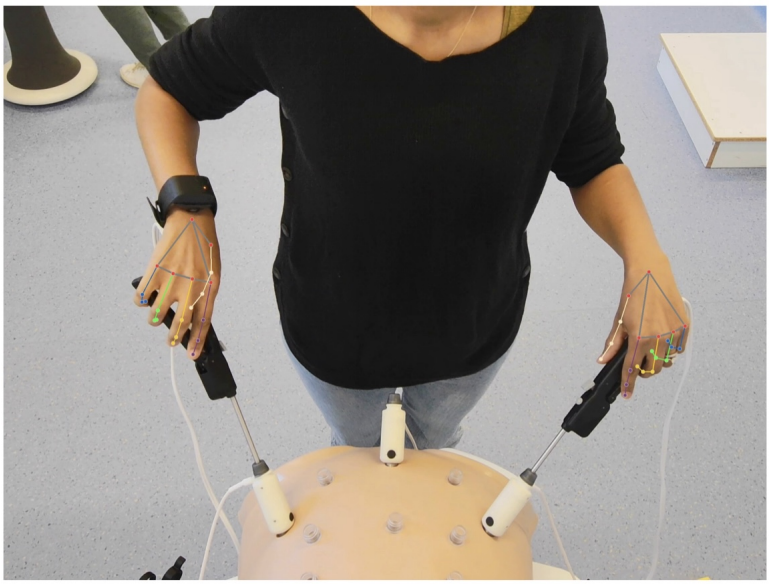


Figure 6.6: A frontal view of a participant performing the needle passing task on the VirtaMed simulator.

6.5.3 Evaluation framework

Circular Cutting Dataset

We employed a subject-independent 6-fold cross-validation strategy to evaluate our framework. This means that data sequences from any one participant, whether they are attending surgeons or surgical residents, are exclusively included in either the training set or the test set, but not both. This evaluation procedure is crucial because hand movement data from the same subjects are likely to exhibit correlations. To ensure a fair distribution of the limited number of attending surgeons across each fold, we generated all possible combinations of two attending surgeons, resulting in six combinations (i.e., six folds). This method guarantees that each surgeon is equally represented in both the training and test sets. For the surgical residents, their data were distributed evenly between the training and test sets across all six folds.

Needle Passing Dataset

For the needle passing dataset, we employed a subject-independent 3-fold cross-validation strategy to evaluate our architecture. We ensured that the number of sequences associated to attending surgeons was fairly distributed between the training and test sets across the three folds.

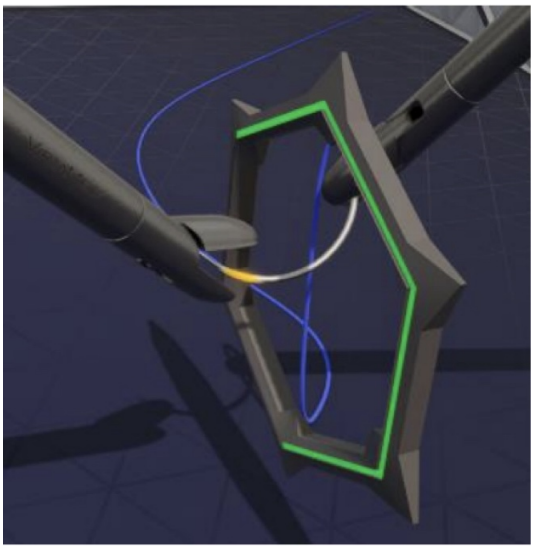


Figure 6.7: Illustration of the needle passing task, captured from the VirtaMed simulator screen.

Evaluation Metrics and Labeling

For both datasets, we used accuracy and weighted average F1-score as evaluation metrics. Accuracy provides a measure of the overall proportion of correctly classified sequences, giving an initial sense of model performance. However, to better handle potential class imbalances between Expert and Novice sequences, we also employed the weighted average F1-score, which combines precision and recall into a single metric that accounts for both false positives and false negatives. This approach ensures that the evaluation fairly reflects the performance across both classes, balancing their contributions to the overall metric. In this chapter, we consider the practitioner's category as an indicator of proficiency, classifying hand skeleton sequences from attending surgeons as Expert sequences and those from surgical residents as Novice sequences. This frames the surgical skill assessment task as a binary classification problem, where the objective is to distinguish between the two classes.

6.5.4 Results

Circular Cutting

We compared our STGFormer framework against eight state-of-the-art models, including both traditional deep learning and advanced graph-based methods. The comparison models comprise deep learning approaches, such as TCN [180], LSTM [181], DeepGRU [182], and Transformer encoder [2], as well as graph-based deep learning models like GCN [3], ST-GCN [156], and ASTGCN [157]. The ST-GCN consists of multiple spatial-temporal convolutional blocks, each of which includes two temporal gated convolution layers and one spatial graph convolution layer in the center. The ASTGCN consists of multiple blocks, each composed of a spatial-temporal attention mechanism and a spatial-temporal convolu-

Table 6.1: Surgical skill assessment on the circular cutting task: comparison with state-of-the-art methods.

Method	Acc	F1
SoCJ-LSTM [179]	80.39	77.55
TCN [180]	80.08	78.25
LSTM [181]	81.21	79.36
DeepGRU [182]	81.42	79.48
Transformer [2]	80.53	78.19
GCN [3]	81.92	80.13
ST-GCN [156]	79.14	79.54
ASTGCN [157]	79.30	79.49
STGFormer (ours)	**83.29**	**81.41**

tion that utilizes graph convolutions to capture spatial patterns and standard convolutions to describe temporal features simultaneously. Additionally, we included the SoCJ-LSTM approach, which uses SoCJ handcrafted features [179] extracted from the hand skeleton. These features are then input into a LSTM model.

Table 6.1 presents the performance of our STGFormer alongside state-of-the-art methods. STGFormer outperforms all other methods on both evaluation metrics, achieving an accuracy of 83.29% and an F1-score of 81.41%. This represents an improvement of 1.37% in accuracy and 1.28% in F1-score over the best-performing comparison model. The SoCJ-LSTM method, which relies on geometric descriptors, achieved the lowest F1-score, highlighting its limitations in capturing the dynamic aspects of surgical skills among sequence of hand landmarks. Even LSTM, which operates directly on raw hand landmarks, performs better but still falls short in comparison to STGFormer.

STGFormer outperformed both GCN and Transformer models, highlighting the effectiveness of integrating GCN with a Transformer encoder for surgical skill assessment. This combination leverages the strengths of GCN in capturing spatiotemporal features and the Transformer encoder in modeling long-range dependencies, demonstrating the powerful synergy of these components in enhancing overall model performance.

Needle Passing

Table 6.2 shows the performance of our STGFormer framework, alongside the other aforementioned state-of-the-art methods on the Needle Passing dataset. We reported the per-

Table 6.2: Performance comparison of STGFormer with state-of-the-art methods on the Needle Passing task across three configurations: left-hand, right-hand, and both-hand skeletons.

	Left		Right		Both	
Method	Acc	F1	Acc	F1	Acc	F1
SoCJ-LSTM [179]	82.53	72.51	84.16	77.66	83.45	76.50
TCN [180]	82.81	75.93	87.28	86.06	88.18	87.13
LSTM [181]	80.75	79.81	87.55	**86.73**	89.07	89.23
DeepGRU [182]	86.62	84.80	87.44	83.56	90.00	88.92
Transformer [2]	**87.25**	84.74	86.76	85.39	88.57	88.79
GCN [3]	84.89	80.98	88.52	85.43	89.87	90.08
ST-GCN [156]	85.11	80.98	86.17	84.78	89.37	88.29
ASTGCN [157]	84.86	80.64	86.33	83.62	90.35	89.81
STGFormer (ours)	87.23	**86.37**	**88.79**	83.69	**91.46**	**91.76**

formance of all models under three configurations: using left-hand skeletons, right-hand skeletons, and combined skeletons from both hands.

For the left-hand skeleton configuration, STGFormer achieved the highest F1-score of 86.37% and an accuracy of 87.23%. Although the Transformer model attained the highest accuracy of 87.25%, its F1-score was 84.74%, falling short of STGFormer by 1.63%. DeepGRU also performed well, with an accuracy of 86.62% and an F1-score of 84.80%. Graph-based methods, such as ST-GCN, also demonstrated competitive performances, achieving an accuracy of 85.11% and an F1-score of 80.98%.

In the right-hand landmarks configuration, STGFormer attained the highest accuracy of 88.79%. However, the highest F1-score was achieved by LSTM at 86.73%, followed closely by TCN with an F1-score of 86.06%. GCN also performed well, with an accuracy of 88.52%, and an F1-score of 85.43%.

When using both hand skeletons, STGFormer outperformed other state-of-the-art models, achieving the highest accuracy of 91.46% and the highest F1-score of 91.76%. LSTM and ASTGCN also demonstrated strong performance, with accuracies of 89.07% and 90.35%, and F1-scores of 89.23% and 89.81%, respectively.

These results consistently demonstrate that STGFormer outperforms other state-of-the-art methods across all configurations (left-hand, right-hand, and combined skeletons). The superior performance observed for most models when using both hand skeletons compared to single-hand configurations highlights the value of incorporating both hands, providing

Table 6.3: Performance comparison of different STGFormer graph configurations on the Needle Passing dataset.

Method	Acc	F1
STGFormer-inter	85.26	82.16
STGFormer-late	86.68	83.13
STGFormer	**91.46**	**91.76**

a richer representation for surgical skill classification.

As with the Circular Cutting Dataset benchmark, the STGFormer outperformed both the GCN and Transformer when using hand data in both hands configuration across all evaluation metrics. Additionally, it achieved a superior F1-score in the left-hand configuration and the highest accuracy in the right-hand configuration. It demonstrates once again that the combination of GCN and Transformer encoder is effective for surgical skill assessment.

Graph Construction Variation

Previously, we presented the performance of the STGFormer framework on the Needle Passing dataset, using sequences of skeleton from both hands. To further explore the impact of hand coordination, we propose comparing STGFormer with two variations that use an alternative graph construction approaches that exclude joint connections between the left and right hands. This comparison aims to highlight the importance of incorporating hand coordination information. To achieve this, we developed two distinct variations: STGFormer-inter and STGFormer-late.

STGFormer-inter employs two independent GCN-Transformer encoder models to process sequences from each hand separately, followed by the fusion of features from both models at an intermediate level, to classify surgical skill. On the other hand, STGFormer-late processes each hand using separate STGFormer models up to the point of prediction, and the final skill classification is achieved by averaging the predictions from both models.

The performance comparisons are presented in Table 6.3. The classical STGFormer, which connects hand joints from both hands, significantly outperforms both variations. Specifically, it surpasses the best-performing comparison by 4.78% in accuracy and 8.63% in F1-score. This superior performance is primarily due to the integration of joint connections between the hands, enabling our framework to effectively capture spatial-temporal interactions and dependencies between the left and right hands. These interactions are critical for accurately assessing surgical skills, as hand coordination plays a crucial role in many surgical tasks. By leveraging this interaction data, the model gains a more comprehensive and nuanced understanding of the surgical procedure. In contrast, the STGFormer-inter

and STGFormer-late variations, which process each hand independently, fail to capture this valuable interaction information, leading to a notable decrease in performance.

6.6 Discussion

This section discusses the implications, limitations, and future directions of our study, highlighting the significance of the findings and the potential impact of the proposed STGFormer framework for surgical skill assessment.

6.6.1 Implications of Findings

The findings of this study demonstrate the potential of using hand skeleton sequences for assessing surgical skills, specifically differentiating between expert and novice practitioners.

Using hand skeleton data for surgical skill assessment offers several advantages over kinematics data and video data. It captures fine motor skills and dexterity by tracking individual finger and joint movements, providing detailed insights into a surgeon's manual technique that robotic kinematics often miss. Unlike kinematics data from surgical robots, which are limited to robotic-assisted surgeries, hand skeleton data is tool-agnostic and can be employed across various surgical tasks, including those performed manually. It provides objective, real-time feedback with less reliance on expensive equipment, enhancing accessibility and reducing costs compared to specialized robotic systems. Unlike video data, hand skeleton data involves simpler processing, is less affected by visual obstructions or lighting variations, and addresses privacy concerns by avoiding the complexities of video analysis. Its compatibility with non-robotic procedures and scalability for assessing multiple trainees simultaneously further solidifies its practicality in diverse training settings, making it an ideal choice for comprehensive and efficient surgical skill assessment.

Furthermore, our proposed STGFormer framework achieved state-of-the-art results on the two surgical simulation datasets. By leveraging the strengths of GCN and Transformer encoder, STGFormer effectively captures spatial-temporal correlations within hand movements, providing accurate assessments of surgical skills. This approach not only demonstrated high accuracy in skill differentiation but also showed robustness across different surgical tasks, reinforcing its adaptability and reliability in various training scenarios.

Our results also confirmed that for all state-of-the-art models, using both hand skeleton sequences significantly improves classification accuracy compared to using a single hand configuration, whether left or right. This underscores the richer representation provided by incorporating both hands in surgical skill assessment. In the case of the STGFormer framework, the way we leverage the connections between both hands has a significant impact on performance. Specifically, in our graph construction, we propose connecting corresponding joints of both hands, which substantially enhances performance.

Subdividing the entire data sequence into non-overlapping 20-second windows allows

for a more detailed analysis of hand movements by capturing granular skill variations within a single task performance. This approach reduces the overall length of data sequences, making the classification task more manageable and computationally efficient, especially when dealing with deep learning models. By focusing on shorter segments, the STGFormer framework can detect subtle differences in hand movement patterns that may indicate the level of expertise, enhancing the model's ability to distinguish between Expert and Novice surgeons effectively. This segmentation strategy also facilitates the augmentation of the dataset, increasing the number of training samples and improving the robustness of the model in identifying variations in skill execution across different tasks.

Lastly, the STGFormer framework's ability to predict expertise levels during surgical simulation tasks highlights its potential utility in real-world educational settings. This tool can provide immediate, objective feedback to practitioners, facilitating the enhancement of surgical simulation training. By automating the assessment process, institutions can efficiently manage larger cohorts of trainees, reduce dependence on subjective evaluations, and maintain consistent training standards. This scalability is particularly advantageous in high-demand specialties where access to expert evaluators is limited, positioning STGFormer as a valuable addition to modern surgical education.

6.6.2 Limitations

Identifying the limitations of our study helps to frame the findings within the context of potential challenges and areas for improvement. It is essential to acknowledge these limitations to guide future research efforts and refine the application of the STGFormer framework.

Firstly, the integration of the GCN and Transformer encoder in our proposed STGFormer framework increases model complexity, leading to higher computational costs, longer training times, and the need for extensive hyperparameter tuning to optimize performance.

Furthermore, the complex architecture of STGFormer poses challenges for interpretability, which is essential in clinical and educational settings where understanding the decision-making process is critical. This lack of transparency may impede the adoption of the model by practitioners who need to trust and comprehend the underlying rationale behind the skill assessments.

Another limitation concerns both collected surgical simulation datasets. The datasets have a limited number of participants, especially attending surgeons (4 and 7), which may not adequately represent the expert population. This small sample size can reduced generalizability of the model. Moreover, each participant performed the task only once for each dataset. This lack of multiple trials per subject limits the ability to capture variability in individual performance, which is crucial for assessing consistency and skill. There is a notable imbalance between the number of attending surgeons and surgical residents in both datasets. This imbalance can introduce bias in the model training, potentially skewing results towards the more represented class (Novice).

Additionally, the binary labeling as either Expert or Novice oversimplifies the assessment by ignoring the spectrum of skill levels within each category. This binary classification may oversimplify the assessment, ignoring subtle gradations in skill that are important for tailored feedback. Moreover, the Expert and Novice labels are based on predefined criteria (e.g., attending surgeons vs. residents), which may not fully account for individual skill variations within those groups. As a result, there may be significant overlap between the most skilled Novices and the less proficient Experts, leading to ambiguous classifications that could affect the model's accuracy and the reliability of its assessments.

Lastly, the subdivision of data sequences into 20-second non-overlapping windows, while enabling a detailed analysis of movement patterns, may result in the loss of contextual information critical to evaluating overall task performance. Important skill-related factors, such as continuity and flow of actions, may not be fully captured within these shorter windows, potentially leading to an incomplete representation of the surgeon's skill level. Moreover, the fixed window length could inadvertently segment key transitions or movements, making it challenging for the model to interpret the complete sequence dynamics necessary for accurate classification.

6.6.3 Future Directions

The promising results of the STGFormer framework for surgical skill assessment suggest several avenues for future research and development. Addressing current limitations and exploring new possibilities will enhance the framework's applicability, robustness, and generalizability in surgical education and beyond.

Future work should focus on collecting larger and more diverse datasets, incorporating a broader range of surgical tasks and participant profiles. Increasing the number of attending surgeons and including multiple trials per participant would capture a wider spectrum of skill variability, enhancing the model's ability to generalize across different individuals and tasks.

Moving beyond binary classification, future studies should aim to develop multi-class or continuous scoring models that better reflect the spectrum of surgical skills. Incorporating a more nuanced evaluation system could enable tailored feedback and more precise skill assessments, accommodating the subtle differences in performance that exist within the categories of Expert and Novice. This approach would better support personalized training and help identify specific areas for improvement.

To facilitate broader adoption in clinical and educational settings, enhancing the interpretability of the STGFormer framework is essential. Future work could explore techniques such as attention visualization or feature importance mapping to make the decision-making process more transparent. This would allow practitioners and educators to understand the factors driving the model's assessments, thereby increasing trust and acceptance of the technology.

The effectiveness of hand skeleton-based assessment relies heavily on the quality of the extracted skeletal data. Future research should explore advanced preprocessing techniques,

such as noise reduction, joint alignment correction, and occlusion handling, to ensure accurate data input in the context of surgical training. Additionally, investigating the use of higher-fidelity sensors or integrating multiple data sources, like wearable sensors alongside video, could further enhance data reliability and assessment precision.

Given the variability in surgical procedures, future research could explore the use of transfer learning to adapt the STGFormer framework to new surgical tasks with minimal retraining. This approach would leverage pre-trained models on existing tasks to accelerate the learning process for novel tasks, making the framework more versatile and reducing the need for extensive new data collection.

6.7 Conclusion

In this chapter, we propose using hand skeleton sequences for the task of surgical skill assessment. Our goal is to differentiate between the hand movement patterns of attending surgeons and those of surgical residents during surgical simulation tasks. To achieve this, we have collected two novel datasets, which included video recording of hand movements from both attending surgeons and surgical residents while executing two surgical simulation tasks: circular cutting and needle passing.

In addition, we introduce STGFormer, a novel deep learning framework specifically tailored for processing hand skeleton sequences. STGFormer combines a GCN and a Transformer encoder to capture spatiotemporal correlations within hand skeleton sequences for surgical skill assessment. The proposed model achieved state-of-the-art performance across both datasets. For the circular cutting dataset, STGFormer surpassed existing methods in the single-hand setting (using the right hand). Additionally, it outperformed previous approaches in both the single-hand (left and right hand) and dual-hand settings for the needle passing dataset.

Moreover, our results on the needle passing dataset demonstrate that employing skeleton sequences from both hands significantly improves performance compared to single-hand configurations, underscoring the richer representations afforded by dual-hand input for surgical skill assessment. Additionally, we show that, for our framework, connecting corresponding joints between both hands during graph construction significantly enhances the results, highlighting the importance of incorporating inter-hand connections at the input level.

The strong performance across both datasets highlights the potential of using hand skeleton sequences for skill level detection. These results suggest the potential of our framework for enhancing surgical simulation training by providing real-time, objective feedback to surgical residents based on their hand movements during simulation tasks.

Nevertheless, despite encouraging results, several challenges remain. The STGFormer model is complex and computationally intensive, requiring substantial resources for both training and inference. This could limit its practical scalability and accessibility. Additionally, there are issues with model interpretability; it is difficult to understand how specific

hand movements influence skill assessments. The datasets used are relatively small and may not generalize well to diverse populations or various surgical contexts, partly due to the limited number of participants, including attending surgeons. Furthermore, the current binary classification of skills as either expert or novice oversimplifies the nuanced spectrum of skill levels.

Future research should focus on addressing these limitations by gathering larger, more diverse datasets, incorporating model interpretability, and adopting a more detailed skill level classification to better support surgical training.

Chapter 7

MGRFormer: A Multimodal Transformer Approach for Surgical Gesture Recognition

Contents

In this chapter, we present novel contributions to the field of surgical gesture recognition. Specifically, we introduce a new multimodal deep learning framework that incorporates an iterative multimodal refinement module design to enhance the fusion of complementary information from kinematic and video modalities during the refinement stage.

In Section 7.1 we discuss the importance of surgical gesture recognition, particularly in the context of surgical education, and outlines the challenges associated with this task. Section 7.2 provides a comprehensive literature review, establishing the necessary background and context for introducing our proposed method. Following this, we introduce our MGR-Former framework in Section 7.3. In Section 7.4, we present and analyze our unimodal and multimodal benchmarks. Section 7.5 provide comparative statistical analysis between surgical gestures performed by attending surgeons and medical students during suturing tasks. Section 7.6 discuss the implications, limitations, and future directions of our work. Finally, Section 7.7 summarizes our contributions and concludes the chapter.

7.1 Introduction

In recent decades, the field of surgery has experienced remarkable advancements driven by cutting-edge technologies and innovative techniques that have revolutionized patient care [183, 184]. As surgical procedures become more complex and precise, the demand for highly skilled surgeons has never been higher.

Mastery of surgical gestures is a critical aspect of surgical training, essential for ensuring patient safety, achieving surgical accuracy and efficiency, and building professional confidence. To develop proficiency in these complex techniques, medical students and surgical trainees often engage in simulation-based training. This type of training allows them to practice and refine their skills in a controlled, risk-free environment. Simulation-based training has become a standard approach in surgical education, offering a safe space for learning and improving surgical skills without compromising patient safety. By incorporating these advanced training methods, the next generation of surgeons will be better prepared to meet the challenges of modern surgery.

The emerging field of surgical gesture recognition holds significant promise for advancing surgical education. Surgical gesture recognition involves classifying automatically the specific actions performed by a practitioner during a surgical procedure. These systems can greatly enhance surgical training by providing detailed, objective feedback on various aspects of a medical student's performance. For instance, gesture recognition systems can identify and track specific gestures, such as "passing the needle through the material" or "cutting the suture" during a suturing task, and offer real-time corrections for any inaccuracies or inefficiencies. Furthermore, these systems can analyze recognized gestures to verify that they are performed in the correct order and calculate performance metrics, such as speed and smoothness, ensuring that students execute each gesture accurately.

By analyzing performance data, gesture recognition systems can tailor training programs to address individual weaknesses, recommending targeted exercises and adapting training modules based on progress. In simulation environments, these systems can create

realistic scenarios by replicating potential complications resulting from incorrect gestures, providing interactive and instructive feedback. This real-time feedback can help students understand the impact of their actions, encouraging them to practice more precise techniques. Additionally, these systems can offer corrective suggestions and allow visualization of the correct techniques, enhancing the interactivity and instructiveness of the simulation.

However, the development of surgical gesture recognition systems presents multiple challenges. Variability in surgical environments, such as differences in lighting conditions, occlusions within the surgical field, and diverse setups and equipment in simulation rooms, can significantly impact the system's performance. Moreover, the variability in how different surgeons execute the same gesture adds complexity, as each may perform the gesture in slightly different ways. To create a robust and accurate system, it is essential to ensure that the dataset encompasses a wide range of surgical simulation procedures, surgeons with varying skill levels, and varying environments conditions. Additionally, acquiring large amounts of annotated surgical data poses a significant challenge due to the need for expert annotation, which is both time-consuming and costly.

To improve the accuracy and robustness of surgical gesture recognition systems, incorporating multiple data modalities, such as kinematic data and video recordings of surgical procedures, can be highly beneficial [29, 30, 31]. These modalities capture distinct yet complementary patterns, offering a more holistic understanding of a surgeon's actions. By integrating these data sources, correlations between hand movements, instrument motions, and visual cues in the video can be more effectively identified. Additionally, a multimodal approach enhances fault tolerance; for example, if unpredictable events like occlusions affect the video, motion sensor data can serve as a backup, ensuring reliable model performance even under challenging conditions.

Nevertheless, integrating multimodal data comes with its own set of challenges. Key issues include differences in data representation and scale, synchronization difficulties, high dimensionality, and potential data loss due to sensor failures or occlusions. An important challenge concerns the effective fusion of different modalities through multimodal learning. This involves determining the optimal stage for data fusion—whether at the early, intermediate, or late stages—and the design of advanced fusion techniques to make the most of the different modalities.

The Transformer architecture [2] has emerged as the predominant choice for a wide range of tasks due to its ability to handle sequential data, including multimodal learning [100, 118], and temporal action segmentation [185, 186, 187]. Inspired by the success of the Transformers in these domains, we introduce MGRFormer, a novel attention-based multimodal framework specifically designed for surgical gesture recognition. MGRFormer leverages the complementary information from kinematic and video modalities at the refinement stage by incorporating a multimodal refinement module. To the best of our knowledge, this is the first work to explore multimodal fusion at the refinement stage. We validate the effectiveness of our approach through extensive experiments on the VTS surgical simulation-based dataset [188].

Additionally, we propose a comparative statistical analysis of the surgical gestures performed by attending surgeons and medical students by calculating multiple performance

metrics. Our objective is to provide a comprehensive solution for enhancing surgical simulation training by integrating surgical gesture recognition with the calculation of performance metrics. This approach will enable a detailed assessment of the proficiency and efficiency of medical students.

The contributions of this chapter are three-fold and can be summarized as follows:

1. We propose a novel multimodal fusion framework that exploits the joint relationship between kinematic and video modalities during the refinement stage.

2. To validate the proposed framework and demonstrate the complementarity between the kinematics and video modalities, we provide both unimodal and multimodal benchmarks.

3. Our MGRFormer significantly outperforms other state-of-the-art methods on the VTS dataset.

7.2 Related Work

In this section, we will review the key techniques and methodologies that form the foundation of our proposed approach. We start by reviewing methods for temporal action segmentation, as our approach is fundamentally based on a method originally developed for this task. Next, we present unimodal and multimodal methods relevant to the task of surgical gesture recognition. Finally, to provide a broader context for our method, we will discuss methods in the field of RGB-D based multimodal gesture recognition.

7.2.1 Temporal Action Segmentation

Temporal action segmentation refers to the localization of individual actions within a video sequence. Traditional methods for identifying actions within video sequences typically involved using a sliding window approach, followed by non-maximum suppression to select the most relevant candidates [189, 190]. Alternative approaches explore the use of Bayesian model [191], Conditional Random Fields [192, 193], and Markov models [194]. Modern approaches for modeling long-range dependencies among actions involve the use of deep neural networks, which encompass a variety of architectures such as Recurrent Neural Networks [195, 196], Temporal Convolutional Networks [197, 198, 199, 200], Graph Neural Networks [201, 202], and recent Transformers [185, 203, 204, 205]. Particularly, the ASFormer [185] has established itself as a state-of-the-art solution for temporal action segmentation. It employs a multi-stage process in which an initial stage generates the initial prediction, followed by subsequent refinement stages responsible for refining and fine-tuning the initial prediction.

While Transformers have achieved remarkable success in temporal segmentation tasks, their application in multimodal contexts remains relatively unexplored. In this chapter, we propose extending the ASFormer for the task of multimodal gesture recognition. We intro-

duce a novel multimodal refinement module, which exploit the complementary information between two different modalities through the use of multiple Transformer decoders.

7.2.2 Surgical Gesture Recognition

Numerous studies proposed the used of kinematics and video data, either independently or in combination, for the task of surgical gesture recognition.

Unimodal

The utilization of robotic kinematics data has been a popular approach due to its precision and the rich set of motion-related features it provides. Early work in this area primarily focused on traditional machine learning techniques. For instance, variants of hidden Markov models [165, 206] have been proposed to classify surgical gestures based on kinematic data.

Recent advancements in deep learning have significantly improved performance, with a rich variety of deep temporal models employed using robotic kinematics data. These models include Convolutional Neural Networks [207], Temporal Convolutional Networks [208, 188], Recurrent Neural Networks [209, 210, 188], and Transformers [211].

On the other hand, video-based gesture recognition has become increasingly popular due to its ability to capture a comprehensive view of the surgical scene. Recent advancements in computer vision and deep learning have significantly improved gesture recognition from video data. For instance, Funke et al. [212] utilized 3D Convolutional Neural Networks, Zhang et al. [213] employed Symmetric Dilated Convolution, and Liu et al. [214] used Deep Reinforcement Learning.

Multimodal

Combining robotic kinematics data with video data can potentially leverage the strengths of both modalities, providing a more comprehensive understanding of surgical gestures. Several studies [193, 215, 29, 216] have reported consistent improvements when combining kinematics and video data compared to the two individual modalities. The integration of these two modalities has been investigated at the input level [216, 215], intermediate level [217, 31], and prediction level [29]. However, a very limited number of studies have explored more complex multimodal approaches. In their paper [29], the authors introduced Fusion-KVE, a novel approach that integrates visual features, kinematics data, and system events. This method employs individual networks for each input modality and then combines their predictions using a weighted voting scheme. Long et al. [30] proposed MRG-Net, an approach that leverages the complementary information between kinematics and visual features using a graph convolutional network. Van Amsterdam et al. [31] introduced MA-TCN, which utilizes multimodal attention mechanisms to weight kinematic and visual features.

Unlike the previously mentioned methods, our approach employs a refinement strategy, specifically a multimodal refinement module that integrates kinematics and video data. Refinement involves enhancing the initial predictions made by a model, typically by applying additional processing or learning techniques. Surgical gesture recognition is a particularly complex task because it requires understanding fine-grained, nuanced movements that can vary significantly between surgeons, procedures, and contexts. A one-stage model, which processes all information in a single pass, often lacks the capacity to fully capture this complexity. By allowing the model to revisit and re-evaluate initial predictions, refinement facilitates incremental adjustments that can better manage variability in the data.

7.2.3 RGB-D based Multimodal Gesture Recognition

The field of multimodal gesture recognition is constantly evolving, with numerous studies exploring various modalities to enhance performance. The integration of RGB and depth data has been extensively investigated for its potential to significantly improve gesture recognition systems. This integration provides crucial spatial context by adding depth information to the RGB data, offering a more comprehensive understanding of gestures. In their study, Hu et al. [218] introduced a novel deep bilinear framework designed to learn time-varying information from multimodal data. Furthermore, for capturing rich modality-temporal patterns, they proposed a novel action feature representation, which encodes the context of RGB-D actions into a tensor structure. Zhou et al. [219] introduced a novel spatial-temporal representation learning framework consisting of decoupled spatial and temporal representation learning networks, denoted as DSN and DTN, respectively, and a recoupling representation learning network denoted as RCM. To effectively exploit multimodal interactions between unimodal branches, they proposed a cross-modal adaptive posterior fusion module, termed CAPF. Furthermore, building upon the previously mentioned work, Zhou et al. [220] introduced a new video data augmentation technique, ShuffleMix, which mask randomly two video pairs along the temporal dimension and then mixes them. They also enhanced the RCM module with a multi-head mechanism that independently generates an attention map for each frame. Furthermore, they introduced a novel cross-modal Complement Feature Catcher (CFCer) for multimodal fusion, aimed at improving the results of late fusion.

7.3 Proposed Approach

In this section, we will introduce our MGRFormer framework, which has been designed for the task of surgical gesture recognition. Figure 7.1 provides an overview of the proposed framework, which consists of three key components: (1) a Kinematics Transformer Encoder, (2) a Vision Transformer Encoder, and (3) a Multimodal Refinement Module.

Initially, kinematic and visual features are extracted using the Kinematics Transformer Encoder and Vision Transformer Encoder, respectively. Both encoders are designed based on the ASFormer Encoder [185]. Subsequently, initial predictions from one chosen modality, along with the extracted features from the other modality, are passed through a series

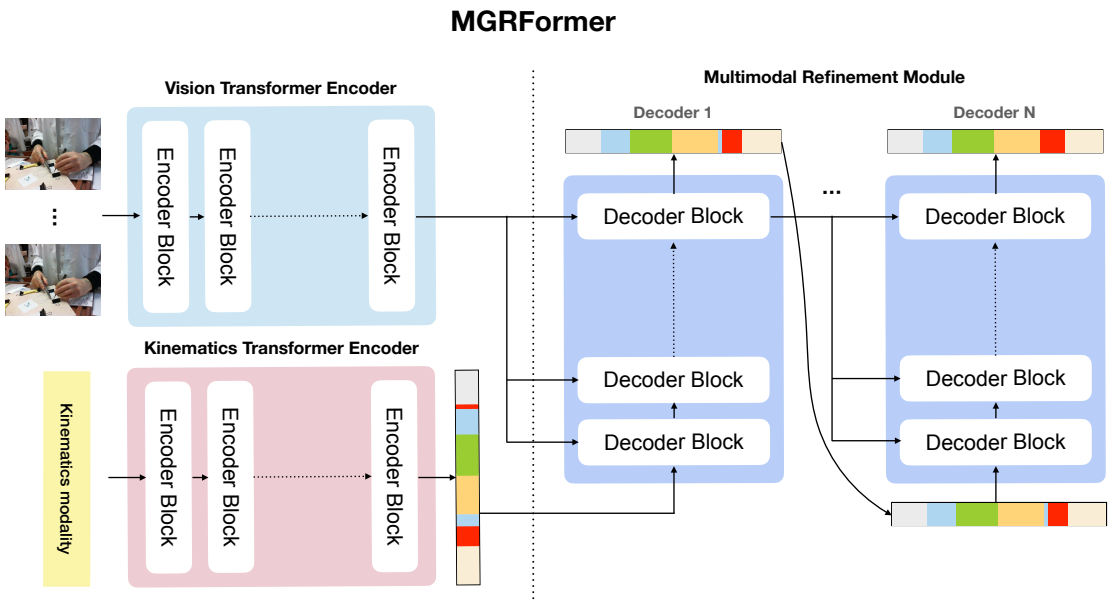


Figure 7.1: Illustration of the MGRFormer framework, consisting of two Unimodal Encoders and a Multimodal Refinement Module for iterative cross-refinement using the output predictions of one modality and the Encoder features of the other modality.

of successive decoders to perform iterative cross-refinement via the proposed multimodal refinement module. This refinement strategy involves progressively enhancing the initial predictions by integrating complementary information from both modalities to achieve more accurate and reliable results. The MGRFormer framework has been designed to predict the probability distributions of surgical gestures for each time step in the data sequence.

7.3.1 Unimodal Transformer Encoder

The first component of our framework includes the Kinematics Transformer Encoder and the Vision Transformer Encoder, which are responsible for extracting kinematic and visual features. The Kinematics Transformer Encoder processes input kinematics data, denoted as x_{kin}, with dimensions $T \times d_{kin}$, while the Vision Transformer Encoder processes visual features, denoted as x_{vis}, with dimensions $T \times d_{vis}$. Here, T represents the sequence length, and d_{kin} and d_{vis} represent the dimensions of kinematics and visual features, respectively.

For the visual features, we used either image features extracted from a pre-trained ResNet-18 [67] or sequence of image features extracted from a pre-trained I3D [221]. The ResNet-18 features have a dimension of 512, while the I3D features have a dimension of 1024. For the I3D feature, we added the RGB and flow predictions.

The initial stage of our framework involves linearly projecting each input feature, x_{kin} and x_{vis}, onto embedding vectors z_{kin} and z_{vis}, respectively. This projection adjusts the dimensionality of the input features. As in the ASFormer model, we have not employed positional encoding, as it has been shown to decrease model performance. Subsequently, each embedding vector is passed through a series of encoder blocks. Finally, a fully connected

layer generates initial predictions for either the kinematics or visual modality, denoted as $\hat{y}_{kin}$ or $\hat{y}_{vis}$. It is important to note that in our framework, only one modality is selected to generate initial predictions.

The encoder is structured as a series of encoder blocks. Each block contains a temporal convolution layer followed by a single-head self-attention layer. A residual connection is incorporated around each of these sub-layers, followed by a ReLU activation function and instance normalization. For more details on the encoder design, refer to [185].

We implemented self-attention using a local window of size w, as proposed in the original ASFormer study. This choice was driven by the need to manage the substantial computational resources required for self-attention calculations in very long videos. The size of the local window increases exponentially with the number of layers ($w = 2^i, i = 1, 2, \ldots$), enabling a smooth transition from a local to a global focus and expanding the receptive field to effectively encompass the entire video sequence. Additionally, we doubled the dilation rate of the temporal convolution layer as the encoder depth increased, ensuring consistency with the self-attention layer.

7.3.2 Multimodal Refinement Module

Motivation

Iterative refinement is a crucial component of modern state-of-the-art methods for temporal action segmentation. In this context, refinement refers to the process of progressively improving a model's initial predictions by using additional layers or stages that incorporate more complex and higher-level contextual information. This iterative process involves adjusting the initial predictions based on the temporal relationships between different actions, allowing the model to better understand and capture the sequence's overall structure. As a result, the refined predictions become more accurate, coherent, and consistent, aligning better with the true temporal dynamics of the action sequences.

As discussed in Section 7.2, multimodal learning has the potential to enhance surgical gesture recognition by the integration of multiple modalities. However, traditional fusion techniques may not be optimal and could result in subpar performance. For instance, the early fusion technique, which combines modalities at the input level, may fail to effectively capture modality-specific patterns and can cause information loss due to discrepancies in data scales, dynamics, or representations. On the other hand, the late fusion mechanism presents its own set of challenges. Since the ASFormer generates multiple prediction outputs, effectively aggregating these outputs across different modalities can be difficult. A straightforward approach might involve using an aggregation function to combine outputs from different modalities at the same level, provided the modalities have an equal number of decoders. However, this method can lead to segmentation errors and may not fully leverage the complementary information available between modalities.

To the best of our knowledge, no studies have explored the development of multimodal fusion techniques specifically at the refinement stage. In this section, we introduce our proposed Multimodal Refinement Module, which employs Transformer decoders to leverage

the complementary information between kinematics and visual modalities during the refinement stage. We will begin by detailing the design of a single decoder and subsequently explain how this design can be extended to incorporate multiple decoders for iterative cross-refinement.

One Decoder

The first decoder takes as input the initial predictions from either the kinematics or the visual modality. These predictions are then passed through a fully-connected layer to adjust their dimensions. The decoder itself consists of a series of decoder blocks, each containing a temporal convolution layer and a cross-attention layer.

In our approach, cross-attention is computed between the encoder features from the visual modality and the output of the preceding decoder block that refines the kinematic initial predictions. Similarly, cross-attention is computed between the encoder features from the kinematics modality and the output of the preceding decoder block that refines the initial visual predictions.

Drawing inspiration from the decoder design in [185], we form the query (Q) and key (K) by concatenating the encoder's output with the previous decoder block's output. The value (V), however, is solely derived from the output of the preceding decoder block. This cross-attention mechanism allows each position in one modality's encoder to attend to all positions in the refinement process of the other modality.

Multiple Decoders

A single decoder might be insufficient to capture the complexity of the data and the intricate relationships between surgical gestures over time. Multiple decoders can increase the model's capacity, enabling it to handle more complex patterns and dependencies.

Expanding from a single decoder to multiple decoders enables further iterative refinement. Each intermediate decoder computes cross-attention between its features and the output predictions from the previous decoder. This iterative process allows for progressively integrating higher-level contextual information and refining predictions at each stage.

Architecture Variations

We proposed different combinations for performing multimodal refinement using the kinematics (k) and video (v) modalities. We denote the process of refining our initial predictions derived from the kinematics modality using the encoder features from the video modality as MGRFormer$_{k\rightarrow v}$. Conversely, MGRFormer$_{v\rightarrow k}$ refers to the situation where we refine our initial predictions from the video modality with the encoder features from the kinematics modality. As we will demonstrate later, a double refinement process can further enhance predictions. For instance, we can refine the predictions from MGRFormer$_{k\rightarrow v}$ using the

kinematics encoder features, resulting in MGRFormer$_{k \rightarrow v+k}$. In Section 7.4, we will report the performance for all possible combinations of single and double refinement between the kinematics and video modalities.

7.3.3 Loss Function

The loss function $\mathcal{L}$ is composed of two parts: a frame-wise classification loss and a smooth loss. The frame-wise classification loss is calculated as the negative log-likelihood of the correct class, and the smooth loss computes the squared error between the probabilities of successive frames. The loss function is defined as follow:

$$\mathcal{L} = \frac{1}{T} \sum_{t} -log(y_{t,\hat{c}}) + \lambda \frac{1}{TC} \sum_{t} \sum_{c} (y_{t-1,c} - y_{t,c})^2$$

Here, $y_{t,\hat{c}}$ denotes the predicted probability for the ground truth label $\hat{c}$ at time t. T represents the total number of points and C the number of distinct surgical gestures. The regularization term λ is fix at 0.60 in our experiments, balancing the classification loss and the smooth loss. The smooth loss aims to encourage consistency in the prediction probabilities between successive frames, which is particularly important for surgical gesture recognition tasks. To train our model, we sum the losses associated with the predictions from both the encoder and decoders.

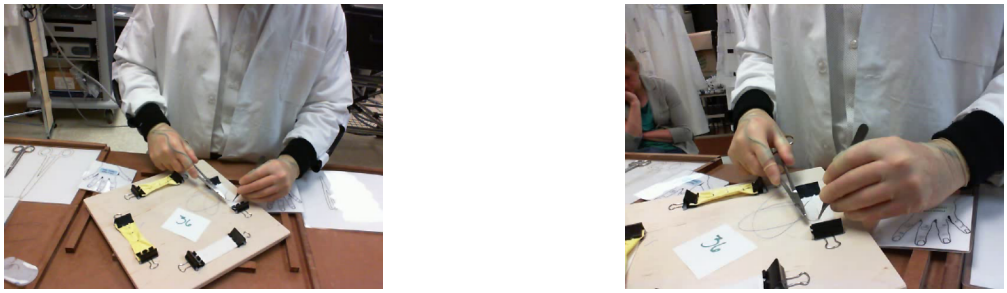
7.3.4 Implementation details

Both Transformer encoders and decoders consist of 10 blocks each. Each input modality is processed by a dedicated Transformer encoder. For single refinement, we employ three decoders. For double refinement, we employ an additional decoder.

As mentioned previously, the input features x_{kin} and x_{vis} are projected onto the embedding vectors z_{kin} and z_{vis}, whose dimension was fix to 128. Following the approach in [185], we applied dropout to the input features of the encoder with a rate of either 0.2 or 0.3, which was chosen through empirical experimentation. In all experiments, we trained our models using the Adam optimizer [113] with a learning rate of 0.0005.

7.4 Experimental Results

In this section, we will start by introducing the VTS dataset used in our experiments. Next, we will present the evaluation metrics and framework employed to assess the performance of the trained models. Lastly, we will discuss the results for the unimodal and multimodal benchmark.



(a) Frontal view (b) Side view

Figure 7.2: Suturing task performed on a tissue sample, observed from two different perspectives. These images have been extracted from the VTS dataset.

7.4.1 Dataset

We conducted our experiments using the Variable Tissue Simulation (VTS) dataset [222], which consists of 24 participants performing a suturing task on two distinct types of tissue simulators. These simulators were designed to represent different material properties: tissue paper was used to simulate friable tissue, while rubber balloons were employed to mimic arterial conditions. Each participant performed the suturing task twice on both simulators, resulting in a total of 96 recorded procedures. The cohort included eleven medical students, one resident, and twelve attending surgeons. One left-handed surgeon was excluded from the study. The duration of each procedure varied from 2 to 6 minutes.

Kinematic data from both hands were captured using electromagnetic motion sensors, while video data were recorded simultaneously by two cameras: a frontal camera focused on the simulation material and a wide-angle camera capturing the surrounding environment, as illustrated in Figure 7.2. Both the sensors and the cameras were synchronized to ensure simultaneous recording.

The suturing exercises were segmented into six distinct gestures:

- G0: “the background gesture"
- G1: “pass the needle through the material"
- G2: “pull the suture"
- G3: “perform an instrumental tie"
- G4: “lay the knot"
- G5: “cut the suture"

Surgical gesture recognition is framed as a multi-class classification task, where the goal is to identify, at each time step of the surgical procedure, one of the six defined surgical gestures.

7.4.2 Evaluation metrics

We evaluated our approach for the task of surgical gesture recognition using two types of metrics: frame-wise and segmentation metrics.

Frame-wise Metrics

- **Accuracy**: This measures the ratio of correctly classified gestures to the total number of frames. It provides a simple and intuitive way to understand the overall performance of the model on a per-frame basis. High accuracy indicates that the model can correctly identify the gesture in each individual frame.
- **Macro F1-score**: This is the average of the F1-scores calculated for each gesture class individually, treating all classes equally. This metric is particularly useful in assessing the model's performance across all gesture classes, especially when the classes are imbalanced.

Segmentation Metrics

- **Segmental Edit Score**: This metric evaluates the structural similarity between the predicted sequence of gestures and the actual sequence by counting the number of operations (insertions, deletions, substitutions) needed to transform the predicted sequence into the actual sequence.
- **Segmental F1-score (F1@k)**: This metric assesses the overlap between predicted and actual gesture segments with varying thresholds (10%, 25%, and 50%). The Segmental F1-score is calculated as follows:

  1. **Define a Matching Criterion**: For each threshold $k\%$, a true positive is counted if the predicted segment overlaps with the ground truth segment by at least $k\%$ of the ground truth segment's length.
  2. **Precision and Recall Calculation**:
     - **Precision**: The ratio of correctly predicted segments (true positives) to the total number of predicted segments (true positives + false positives).
     - **Recall**: The ratio of correctly predicted segments (true positives) to the total number of actual segments (true positives + false negatives).
  3. **F1-Score Computation**: The F1 Score is the harmonic mean of precision and recall, given by:

  $$F1@k = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

  High segmental F1-score indicate that the model not only recognizes gestures correctly but also precisely identifies when each gesture starts and ends.

By using both frame-wise and segmentation metrics, we can comprehensively evaluate our approach. Frame-wise metrics provide a detailed view of the model's performance at the granular level, ensuring that each frame is correctly classified. Segmentation metrics, in contrast, offer insights into the temporal structure and boundary accuracy of the gesture sequences. Together, these metrics ensure a thorough and robust evaluation of the model's performance in recognizing surgical gestures.

Method	Modality	Features	Acc	F1-Macro	Edit	F1@{10,25,50}		
LSTM [188]	kin	✗	81.26	77.05	84.69	88.07	83.69	68.13
GRU [188]	kin	✗	82.23	78.20	84.94	88.01	83.82	68.86
MS-TCN++ [188, 200]	kin	✗	82.40	78.92	86.30	89.30	85.79	71.12
ASFormer [185]	kin	✗	82.66	79.46	88.65	91.36	87.68	72.55
MS-TCN++ [200]	frontal	ResNet-18	77.84	73.34	77.80	81.36	78.21	63.87
MS-TCN++ [200]	frontal	I3D	82.85	78.85	86.33	89.98	87.39	74.33
ASFormer [185]	frontal	ResNet-18	79.25	75.20	84.17	87.16	83.86	69.00
ASFormer [185]	frontal	I3D	82.72	78.90	88.28	91.28	88.35	73.80
MS-TCN++ [200]	side	ResNet-18	84.45	81.71	82.35	87.01	85.32	77.01
MS-TCN++ [200]	side	I3D	86.83	84.14	86.68	90.85	89.83	82.48
ASFormer [185]	side	ResNet-18	85.44	82.87	86.26	90.44	88.98	80.41
ASFormer [185]	side	I3D	87.43	85.29	89.24	92.89	91.61	85.05

Table 7.1: Unimodal surgical gesture recognition. The terms "kin", "frontal", and "side" refer to the specific modalities employed: kinematics data, frontal video, and side view video, respectively.

7.4.3 Evaluation framework

Following prior works [188], we employed a subject-independent 5-fold cross-validation strategy to train all our models. In each fold, the dataset was split into training, validation, and test sets, following the methodology described in [188]. For each evaluation metric, we reported the mean across all folds.

7.4.4 Results

We conducted both unimodal and multimodal benchmarks using kinematic data and video from frontal and side views. For each video view, we employed features extracted using ResNet-18 and I3D to evaluate MGRFormer's capability to effectively handle both image and video features, alongside kinematics data.

Unimodal

In Table 7.1, we present the performance of the ASFormer model, alongside results from several state-of-the-art methods, across three modalities: kinematics, frontal-view, and side-view video. The ASFormer consistently outperforms other methods across all input modal-

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
Fusion-KV [29]	81.94	77.28	83.33	87.21	83.18	68.25
MGR-Net [30]	77.70	73.87	81.49	85.08	80.64	62.17
MA-TCN [31]	79.91	75.64	82.02	86.21	82.32	66.38
MS-TCN++ (early)	82.01	79.07	82.54	86.65	83.97	71.26
MS-TCN++ (late)	82.77	79.56	86.69	89.32	85.52	70.84
ASFormer (early)	81.15	77.35	85.66	88.59	86.01	72.25
ASFormer (late)	81.85	77.82	84.04	88.12	85.02	71.58
MGRFormer $v \rightarrow k$	82.80	79.29	88.06	91.55	88.50	73.61
MGRFormer $k \rightarrow v$	83.85	80.35	88.34	91.28	88.12	74.71
MGRFormer $v \rightarrow v + k$	80.66	76.69	84.67	87.93	85.31	70.39
MGRFormer $k \rightarrow k + v$	83.81	80.47	88.22	91.81	89.14	76.28
MGRFormer $v \rightarrow k + v$	82.07	78.46	87.22	90.40	87.43	73.40
MGRFormer $k \rightarrow v + k$	**84.05**	**80.66**	**89.14**	**92.30**	**89.80**	**76.40**

Table 7.2: Multimodal surgical gesture recognition: kinematics + frontal view (ResNet-18 features). Regarding the notation for MGRFormer, the prediction derived from the modality on the left side of the arrow is refined using the modalities on the right side. For instance, MGRFormer$_{k \rightarrow v+k}$ denotes the process where the kinematics prediction is first refined with video features, followed by a subsequent refinement using kinematics features.

ities. Regarding the kinematics modality, we observed significant improvements of at least 0.54%, 2.35%, and 2.06% in terms of macro F1-score, Edit score, and F1@10, respectively. For both frontal and side view modalities with ResNet-18 features, the ASFormer consistently outperformed the MS-TCN++ across all types of extracted features. Specifically, for the side view modality, the ASFormer surpassed the MS-TCN++ by 1.16%, 3.91%, and 3.43% in terms of the macro F1-score, Edit score, and F1@10, respectively.

The ASFormer exhibits the best performances for all evaluation metrics by using the side view modality combined with I3D features. Conversely, the ASFormer shows the poorest performance when employing the frontal view modality with ResNet-18 features. Furthermore, for both the frontal and side view modalities, we observe that using I3D features yields superior performance compared to ResNet-18 features. This enhancement can be attributed to the fact that I3D features are better suited for capturing temporal correlations among adjacent frames. In contrast, ResNet-18 features are extracted from individual images, neglecting the contextual information provided by neighboring frames.

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
Fusion-KV [29]	81.82	77.70	84.42	87.62	83.32	68.69
MGR-Net [30]	78.88	75.56	81.63	85.93	82.62	64.79
MA-TCN [31]	83.15	80.04	84.50	88.38	85.98	73.33
MS-TCN++ (early)	85.17	83.21	84.77	89.22	88.01	80.05
MS-TCN++ (late)	86.81	83.90	82.83	88.00	86.20	78.35
ASFormer (early)	85.76	83.42	86.93	90.76	89.26	80.80
ASFormer (late)	85.53	83.02	85.69	89.67	88.10	80.11
MGRFormer $v \rightarrow k$	85.95	83.47	89.24	92.78	91.16	81.58
MGRFormer $k \rightarrow v$	87.40	85.17	89.53	93.08	**91.78**	84.02
MGRFormer $v \rightarrow v + k$	84.97	82.16	86.51	90.19	88.81	79.62
MGRFormer $k \rightarrow k + v$	86.75	84.34	88.58	91.91	90.37	82.68
MGRFormer $v \rightarrow k + v$	84.81	82.16	86.70	90.43	88.98	80.17
MGRFormer $k \rightarrow v + k$	**87.61**	**85.47**	**89.74**	**93.40**	**91.77**	**85.12**

Table 7.3: Multimodal surgical gesture recognition: kinematics + side view (ResNet-18 features).

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
Vision Encoder	85.49	83.02	81.48	86.73	85.00	76.34
Kinematics Encoder	83.64	80.36	83.09	87.41	83.62	69.12
One Decoder	86.09	83.47	86.54	90.09	88.88	80.36
Two Decoders	86.15	83.75	87.91	91.51	89.97	82.19
Three Decoders (ours)	**87.40**	**85.17**	**89.53**	**93.08**	**91.78**	**84.02**
Four Decoders	85.26	82.71	87.52	91.14	89.79	81.19

Table 7.4: Comparative results from varying the number of decoders in MGRFormer$_{k \rightarrow v}$, using kinematics data and side view video with ResNet-18 features. The performance when using only vision and kinematics encoders are also included.

Multimodal

We present the results regarding the fusion of the kinematics and video modalities in Tables 7.2, 7.3, 7.5, and 7.6. We benchmarked our method against several state-of-the-art multimodal methods that integrate kinematics with frontal and side view videos, using ResNet-18 features. These techniques include Fusion-KV [29], MGR-Net [30], and MA-TCN [31]. Specifically for MGR-Net, we re-implemented the entire framework excluding the LSTM module, as its inclusion leads to lower performance. Our comparison also featured MS-TCN++ [200], a state-of-the-art approach in action segmentation, which employs an iterative refinement. It should be noted that this particular refinement is different from the one presented in our work. Furthermore, we tested MS-TCN++ under two classical multimodal fusion settings: early and late fusion. The results of these comparisons are detailed in Tables 7.2 and 7.3. Our MGRFormer outperformed all the aforementioned state-of-the-art methods by a large margin in merging kinematics with both video perspectives. Specifically, for the side view modality, MGRFormer$_{k\rightarrow v+k}$ exceeded the performance of Fusion-KV, MGR-Net, and MA-TCN by minimum margins of 5.43%, 5.24%, and 5.02%, respectively, in terms of macro F1-score, Edit score, and F1@10. It also surpassed both the early and late fusion variants of MS-TCN++, but to a lesser extent. We observed that both multimodal versions of MS-TCN++ outperformed the other three baseline models. This enhancement is likely due to MS-TCN++'s iterative refinement module, which boosts the network's accuracy by repeatedly refining gesture segment predictions.

To demonstrate the effectiveness of the multimodal refinement module, we conducted an ablation study on the number of decoders in MGRFormer$_{k\rightarrow v}$, where we fused kinematics and side view video with ResNet-18 features. As shown in Table 7.4, it was found that selecting three decoders for iterative cross-refinement yielded the best performance across all metrics. It was observed that adding another decoder beyond three did not lead to further improvement, while it did add more complexity to the overall model. Furthermore, we can observe that using at least one decoder significantly improves performance compared to both the vision and kinematics encoders, which demonstrates the utility of the cross-refinement module.

The MGRFormer architecture consistently outperformed each modality when used individually. When integrating kinematics data and frontal view video features extracted using ResNet-18, the MGRFormer$_{k\rightarrow v+k}$ model significantly outperformed each input modality when used separately, as demonstrated in Table 7.2. We observed enhancements of 1.20%, 0.49%, and 0.94% in terms of macro F1-score, Edit score, and F1@10, compared to the unimodal ASFormer trained on the kinematics data. Similarly, improvements of 5.46%, 4.97%, and 5.14% were noted in comparison to the ASFormer trained with ResNet-18 features from the frontal view modality. As for the fusion of kinematics data and the side view video with ResNet-18 features, we observed significant improvements of at least 2.60% in macro F1-score, 1.09% in Edit score, and 2.04% in F1@10, compared to the best results obtained from each of the two individual modalities (see Table 7.3). Similar improvements were observed with I3D features, as shown in Tables 7.5 and 7.6.

When comparing the results of the ResNet-18 and I3D features in combination with the kinematics modality, the MGRFormer exhibits slightly superior performance when utilizing

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
ASFormer (early)	83.62	**80.15**	88.09	91.66	89.32	**76.86**
ASFormer (late)	**83.73**	80.13	86.73	90.20	87.61	74.92
MGRFormer $v \rightarrow k$	82.74	79.21	88.00	91.30	88.69	74.92
MGRFormer $k \rightarrow v$	83.12	79.77	**89.53**	**92.44**	**89.48**	74.79
MGRFormer $v \rightarrow v+k$	82.60	79.07	87.85	91.20	88.41	75.45
MGRFormer $k \rightarrow k+v$	83.21	79.29	87.47	90.90	87.86	74.09
MGRFormer $v \rightarrow k+v$	83.12	79.57	88.60	91.79	89.05	75.81
MGRFormer $k \rightarrow v+k$	82.95	79.44	88.36	92.01	89.30	74.14

Table 7.5: Multimodal surgical gesture recognition: kinematics + frontal view (I3D features).

the I3D features in regard of the side view modality (see Table 7.3 and 7.6). However, the opposite effect can be observed when employing the frontal view modality (see Table 7.2 and 7.5).

To highlight the effectiveness of our MGRFormer framework compared to conventional fusion techniques, we performed a comparative analysis against traditional multimodal fusion methods, including early fusion and late fusion. Specifically, ASFormer (early) concatenates the kinematics and video modalities at the input level, while ASFormer (late) adds the predictions from both the encoders and decoders of the different modalities. For this particular case, it is worth noting that both ASFormer instances for each input modality must have the same number of encoders and decoders to add the predictions from both modalities of the same stage. By combining kinematics and side-view modalities with ResNet-18 features, MGRFormer achieved significant improvements over ASFormer (late). Specifically, we achieved a 2.45% increase in F1-score, a 4.05% improvement in the Edit score, and a 3.73% enhancement in F1@10 with the configuration MGRFormer$_{k \rightarrow v+k}$ (see Table 7.3). Similarly, when compared to ASFormer (early), MGRFormer$_{k \rightarrow v+k}$ demonstrated improvements of 2.05%, 2.81%, and 2.64% in F1-score, Edit score, and F1@10, respectively.

For the frontal view modality, as shown in Table 7.2, MGRFormer$_{k \rightarrow v+k}$ demonstrates an improvement of 2.84%, 5.10%, and 4.18% in F1-score, Edit score, and F1@10, respectively, compared to ASFormer (late). Furthermore, we observe improvements of 3.31%, 3.48%, and 3.71% in these metrics compared to ASFormer (early). When using I3D features with the side view modality, there is an improvement over both baselines, albeit to a lesser extent, as depicted in Table 7.6. However, for the frontal view modality with I3D features, Table 7.5 shows that MGRFormer$_{k \rightarrow v+k}$ achieves only marginal improvements in Edit score, F1@10, and F1@25 compared to ASFormer (early) and ASFormer (late).

These results demonstrate the superiority of our proposed method over traditional multimodal approaches. Despite the notable performance gains, the complexity of our MGRFormer model remains comparable to that of ASFormer (early). With a single cross-

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
ASFormer (early)	87.62	85.20	88.55	92.23	91.16	84.09
ASFormer (late)	87.68	85.13	86.62	90.83	89.69	82.89
MGRFormer $v \rightarrow k$	86.90	84.57	88.26	91.91	90.82	83.86
MGRFormer $k \rightarrow v$	**88.39**	**86.03**	89.55	93.46	92.38	**86.29**
MGRFormer $v \rightarrow v + k$	87.24	84.47	89.11	92.36	91.23	84.47
MGRFormer $k \rightarrow k + v$	87.47	85.31	87.81	91.85	90.32	83.46
MGRFormer $v \rightarrow k + v$	87.44	85.09	89.54	92.61	91.51	84.93
MGRFormer $k \rightarrow v + k$	88.10	85.89	**89.91**	**93.51**	**92.40**	85.66

Table 7.6: Multimodal surgical gesture recognition: kinematics + side view (I3D features).

refinement stage, MGRFormer requires only one additional encoder, while maintaining the same number of decoders as ASFormer (early). The double cross-refinement version of MGRFormer introduces only one extra decoder, resulting in a modest increase in complexity, remaining only slightly more complex than in the single cross-refinement setting. In contrast, compared to ASFormer (late), our method is significantly more efficient, requiring only half the number of decoders in regard of the single cross-refinement setting. ASFormer (late) demands training two separate models, each with one encoder and multiple decoders.

Finally, regarding the various settings associated with our MGRFormer framework, we observe that the one-stage refinement model, MGRFormer$_{k \rightarrow v}$, consistently outperforms MGRFormer$_{v \rightarrow k}$ across all combinations of kinematic data and video views when evaluated with ResNet-18 and I3D features, as shown in Tables 7.2, 7.3, 7.5, and 7.6. For instance, Table 7.6 demonstrates that MGRFormer$_{k \rightarrow v}$ surpasses MGRFormer$_{v \rightarrow k}$ in terms of F1-score, Edit score, and F1@10 by 1.46%, 1.29%, and 1.55%, respectively. These results highlight the superior performance of our framework in leveraging video encoder features to refine initial kinematic predictions compared to the inverse approach.

The advantage of MGRFormer$_{k \rightarrow v}$ can be attributed to the richer spatiotemporal context provided by video data, which is critical for iterative refinement. Surgical gestures, characterized by intricate, fine-grained movements and interactions with various tools and tissues, are more discernible in video data. This modality captures not only the detailed visual context of the surgical site, including tool-tissue interactions and surgeon hand movements, but also the subtleties necessary for accurately identifying gestures. In contrast, while kinematic data is valuable, it lacks the visual nuances essential for distinguishing between closely related gestures and focuses primarily on motion trajectories.

This observation is further supported by the findings in Table 7.4, where training the Transformer encoder with side-view video data outperforms training with kinematic data across several metrics, including accuracy, F1-score, F1@25, and F1@50. These results demonstrate the superior contextual robustness of video data for gesture segmentation.

Figure 7.3: Color-coded illustration of surgical gesture recognition on the VTS dataset, comparing ground truth with MGRFormer$_{k\to v}$ predictions, trained using kinematics data and I3D features from the side view.

However, kinematic data, which surpasses side-view video in terms of Edit score and F1@10, can still enhance video-based predictions through our proposed multimodal refinement module—though to a lesser extent than when fusing kinematic predictions with video features.

Furthermore, we reported results for all possible combinations of double refinements involving kinematics and video modalities. As shown in Tables 7.2, 7.3, and 7.6, MGRFormer$_{k\to v+k}$ consistently outperformed the other combinations across each evaluation metric. This finding aligns with expectations, as MGRFormer$_{k\to v}$ achieved the best results for one-stage refinement. When comparing single and double refinements, we observed that MGRFormer$_{k\to v+k}$ is more effective than MGRFormer$_{k\to v}$ when integrating kinematics and video data from both views using ResNet-18 features, as evidenced in Tables 7.2 and 7.3. Specifically, for the I3D features, MGRFormer$_{k\to v+k}$ achieves superior results compared to MGRFormer$_{k\to v}$ in terms of Edit score, F1@10, and F1@50 when fusing kinematics with side view video, as shown in Table 7.6. In contrast, for the fusion of kinematics data with frontal view video, MGRFormer$_{k\to v}$ outperforms MGRFormer$_{k\to v+k}$ across all six evaluation metrics (refer to Table 7.5).

Figure 7.3 presents a visualization of the predictions of our proposed MGRFormer$_{k\to v}$ framework, which integrates kinematic data with I3D features extracted from the side-view video, compared against ground truth for a sequence in the testing set. This visualization highlights the temporal consistency of surgical gesture predictions achieved by leveraging multimodal data with the MGRFormer model.

7.5 Surgical Gesture Analysis

In this section, we present a statistical analysis of the surgical gestures performed by attending surgeons and medical students during suturing tasks, using the kinematics data from the VTS dataset. Our goal is to assess the proficiency and efficiency of both type of practioners across the different surgical gestures performed. By combining our proposed MGRFormer model for surgical gesture prediction during suturing tasks with the calculation of relevant performance metrics, we aim to provide objective feedback at the level of surgical gestures.

This analysis will be particularly valuable for comparing the proficiency of learners to that of experienced surgeons across different surgical gestures, thereby enabling the development of targeted training programs for medical students and less experienced surgeons. Additionally, these performance metrics will support the continuous monitoring of medical students' progress over time.

The performance metrics we will compute include gesture duration, gesture frequency, path length, gesture speed, gesture acceleration, gesture smoothness, and gesture curvature, which are well-established metrics proven to be effective for surgical skill analysis [222, 166, 223]. These metrics can provide a comprehensive assessment of both the efficiency and precision of surgical movements.

In the following, we will introduce each of the aforementioned performance metric and then provide an analysis of the results associated with each metric.

7.5.1 Surgical Performance Metrics

The performance metrics will be derived from the barycenter position of each hand, calculated using the spatial coordinates from the three sensors positioned on each hand. These metrics will be computed for each defined surgical gesture.

Gesture Completion Time

Description: The total time taken to perform a surgical gesture.
Calculation: Measure the time from the beginning to the end of the surgical gesture.
Equation:

$$T = t_{\text{end}} - t_{\text{start}} \tag{7.1}$$

Where T denotes the completion time of the gesture perform, with t_{start} and t_{end} representing the start and end times of the gesture, respectively.

Gesture Frequency

Description: The number of times each specific gesture is performed within the suturing task.
Calculation: Count the occurrences of each surgical gesture (G0 to G5) within the suturing procedure.
Equation:

$$F_g = \sum_{i=1}^{N} I(g_i = G) \tag{7.2}$$

Where F_g is the frequency of surgical gesture G, N is the total number of perform surgical gestures in the suturing procedure, and I is the indicator function that equals 1 when gesture g_i is G, and 0 otherwise.

Path Length

Description: The total distance traveled by the hand during the execution of a surgical gesture.
Calculation: Compute the distance between consecutive barycenter positions of the hand and sum these distances.
Equation:

$$L = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2} \tag{7.3}$$

Where L is the path length, (x_i, y_i, z_i) and $(x_{i+1}, y_{i+1}, z_{i+1})$ are consecutive barycenter of the hand positions, and N is the number of data points.

Gesture Speed

Description: The average speed of the hand for a given surgical gesture.
Calculation: Divide the total path length by the total duration of the gesture.
Equation:

$$S = \frac{L}{T} \tag{7.4}$$

Where S is the average speed of execution, L is the path length, and T is the total duration of the gesture.

Gesture Acceleration

Description: The average absolute rate of change in the speed of the hand.
Calculation: Compute the average acceleration over the duration of the procedure.
Equation:

$$a_i = \frac{|v_{i+1} - v_i|}{t_{i+1} - t_i} \tag{7.5}$$

$$A = \frac{1}{N-2} \sum_{i=1}^{N-2} a_i \tag{7.6}$$

Where N is the total number of gestures performed during the suturing task, a_i is the absolute acceleration at time t_i, and A is the average acceleration.

Gesture Smoothness

Description: The fluidity of the gesture, often measured by the jerk (rate of change of acceleration).
Calculation: Compute the standard deviation of the absolute jerk values over time. This

provides an indication of the variability in the smoothness of the gesture.
Equation:

$$j_i = \frac{|a_{i+1} - a_i|}{t_{i+1} - t_i} \tag{7.7}$$

Where j_i is absolute jerk at time t_i.

$$J_{\text{SD}} = \text{SD}(j) \tag{7.8}$$

Where J_{SD} is the standard deviation of the absolute jerk values j.

Curvature of Path

Description: How sharply the barycenter's path deviates from a straight line.
Calculation: Compute the curvature using consecutive position vectors.
Equation:

$$\text{Curvature} = \frac{|\mathbf{v1} \times \mathbf{v2}|}{|\mathbf{v1}| \cdot |\mathbf{v2}|} \tag{7.9}$$

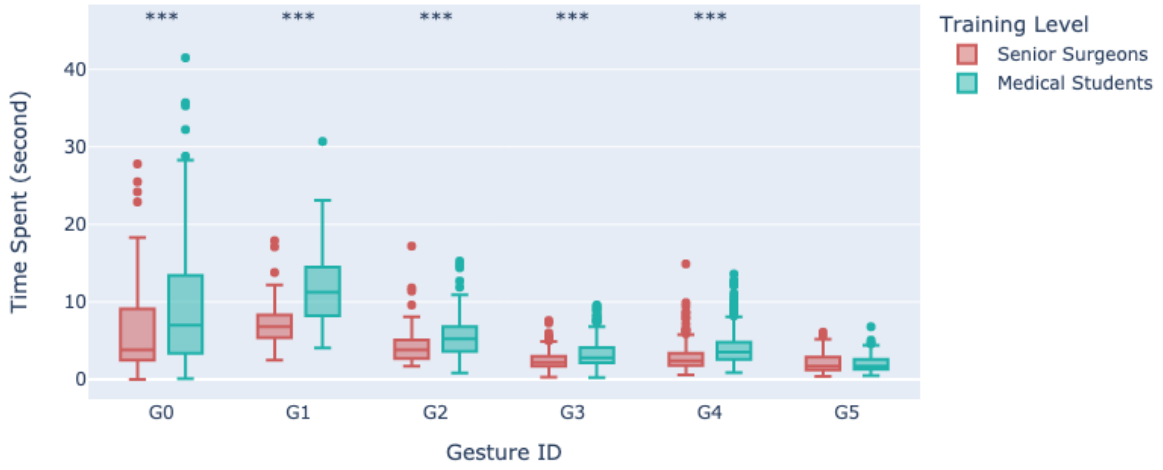
Where:

- $\mathbf{v1}$ is the vector from the barycenter position at time t_i to the barycenter position at time t_{i+1}:
$$\mathbf{v1} = (x_{i+1} - x_i, y_{i+1} - y_i, z_{i+1} - z_i) \tag{7.10}$$
- $\mathbf{v2}$ is the vector from the barycenter position at time t_{i+1} to the barycenter position at time t_{i+2}:
$$\mathbf{v2} = (x_{i+2} - x_{i+1}, y_{i+2} - y_{i+1}, z_{i+2} - z_{i+1}) \tag{7.11}$$
- $\mathbf{v1} \times \mathbf{v2}$ denotes the cross product of $\mathbf{v1}$ and $\mathbf{v2}$, which results in a vector perpendicular to the plane formed by $\mathbf{v1}$ and $\mathbf{v2}$.
- $|\mathbf{v1}|$ and $|\mathbf{v2}|$ are the magnitudes of vectors $\mathbf{v1}$ and $\mathbf{v2}$, respectively.

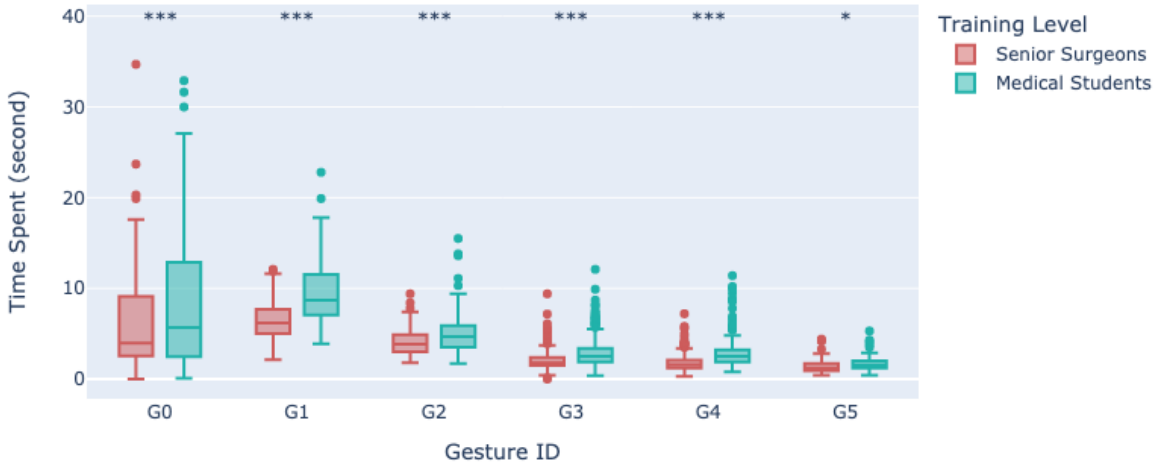
7.5.2 Performance Analysis

Performance metrics will be calculated separately for attending surgeons and medical students for each of the previously defined surgical gestures. For gesture completion time and frequency metrics, results will be reported for both the tissue and balloon simulators. For all other performance metrics, results will be presented for both hands but only for the tissue simulator, as including performance data for both simulators would be redundant and overly detailed.

For each performance metric, we present a box plot for each practionner type and across all gesture. Furthermore, the differences in means between the two groups (attending surgeons and medical students) were analyzed using an independent two-sample t-test for



(a) Tissue Simulator

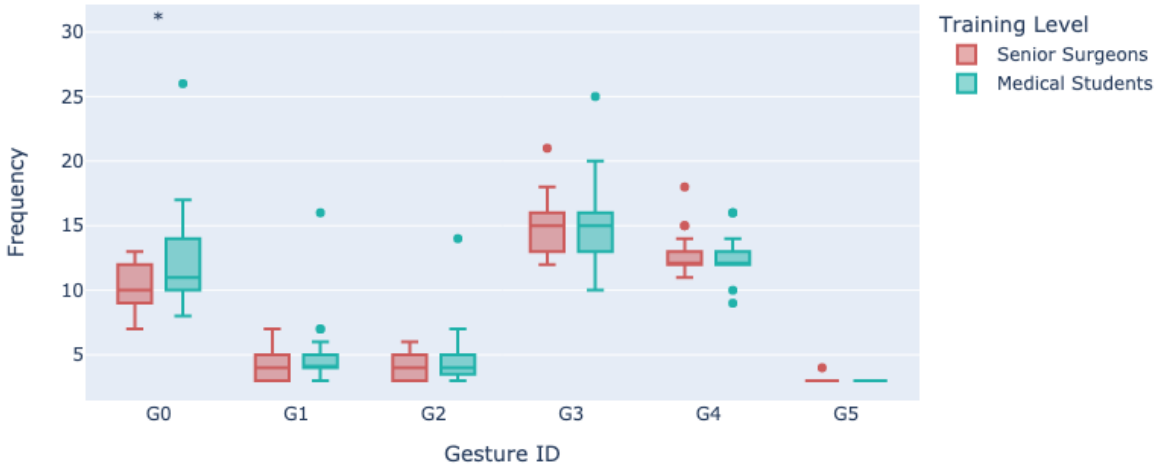


(b) Balloon Simulator

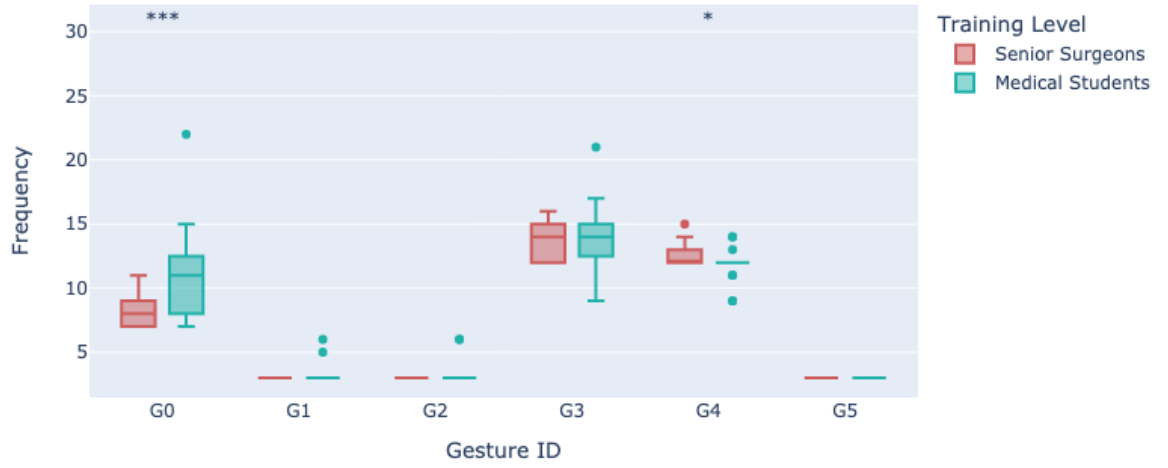
Figure 7.4: Box plots comparing the time spent (in seconds) to complete various surgical gestures (G0 to G5) during suture procedures by attending surgeons and medical students. The procedures were performed using two different types of simulators. Significance levels are indicated as follows: * p-value < 0.05, ** p-value < 0.01, and *** p-value < 0.001.

all gestures. The statistical significance of the differences is indicated in the box plots as follows: * p-value < 0.05, ** p-value < 0.01, and *** p-value < 0.001. This test was chosen because it allows for comparing the means of two independent groups to determine if there is a statistically significant difference between them.

To enhance the clarity of the box plots, especially given the varying spread and the presence of outliers across different gestures, the plots display data up to the 99th percentile of the values across all gestures. The 1% of the largest values were omitted to prevent extreme outliers from distorting the visualization, making the central distribution of data more interpretable.



(a) Tissue Simulator

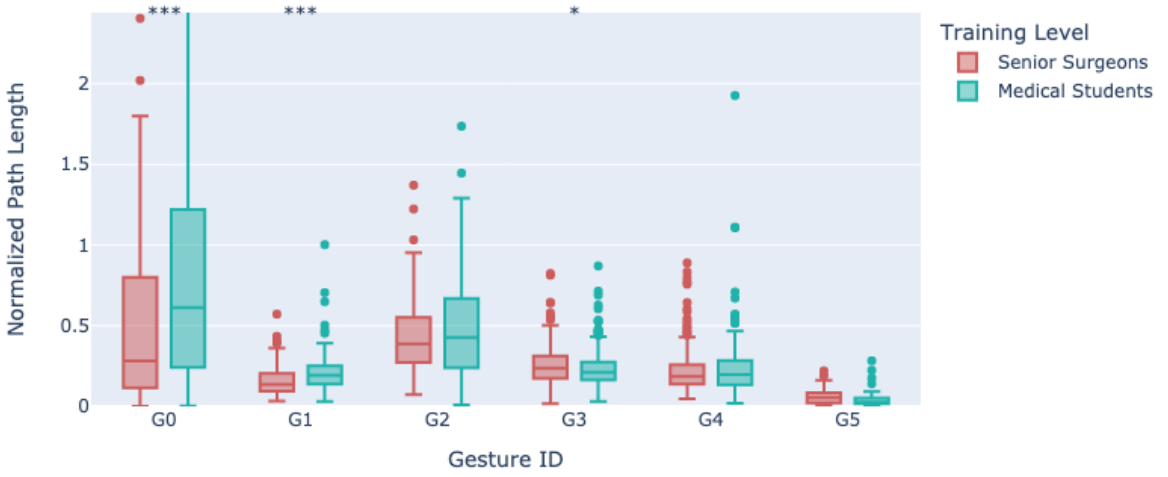


(b) Balloon Simulator

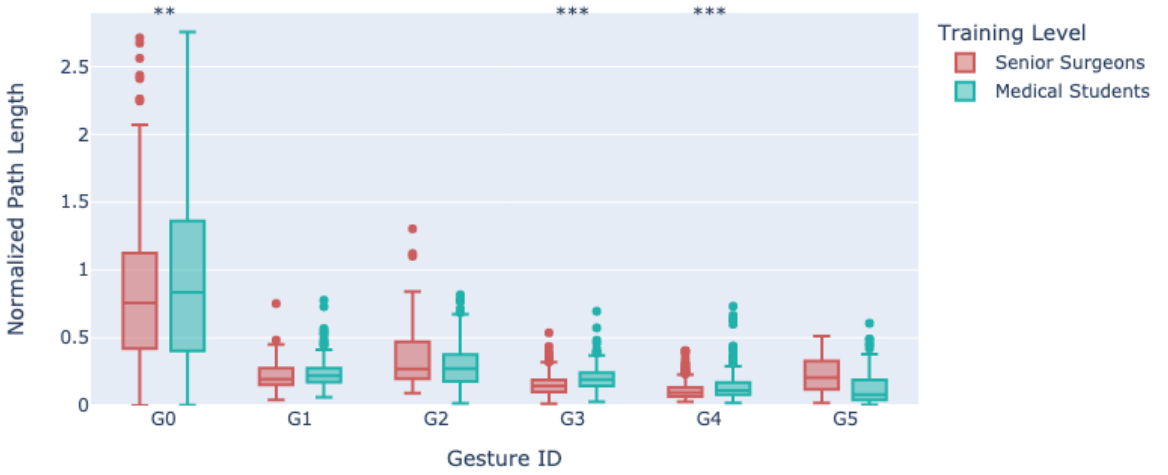
Figure 7.5: Box plots showing the frequencies of different surgical gestures (G0 to G5) during suture procedures performed by attending surgeons and medical students using two types of simulators.

Gesture Completion Time: The box plots in Figure 7.4 illustrate the time spent on each surgical gesture by both attending surgeons and medical students on both tissue simulators. Medical students consistently took longer compared to attending surgeons across nearly all surgical gestures and tissue simulators. For instance, students spent significantly more time on gestures such as passing the needle through the material (G1) and pulling the suture (G2).

Gesture Frequency: Frequencies associated with each surgical gesture, reported in Figure 7.5, reveal notable differences between attending surgeons and medical students. Medical students demonstrated a significantly higher frequency of the background gesture (G0) across both simulators, suggesting a higher level of correction or adjustment during the suturing procedure, indicative of their relative inexperience. Conversely, attending surgeons, due to their extensive experience, performed fewer adjustments and pauses.



(a) Left Hand

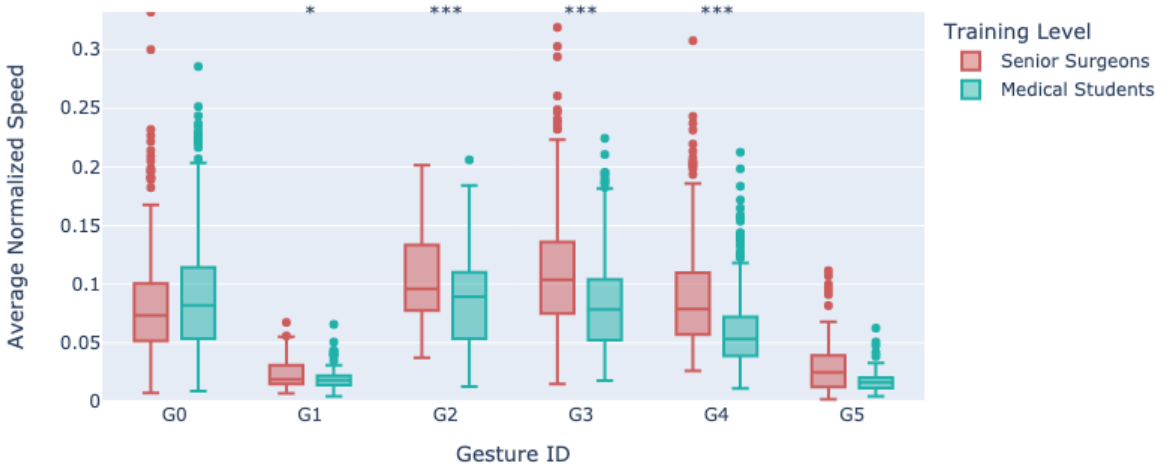


(b) Right Hand

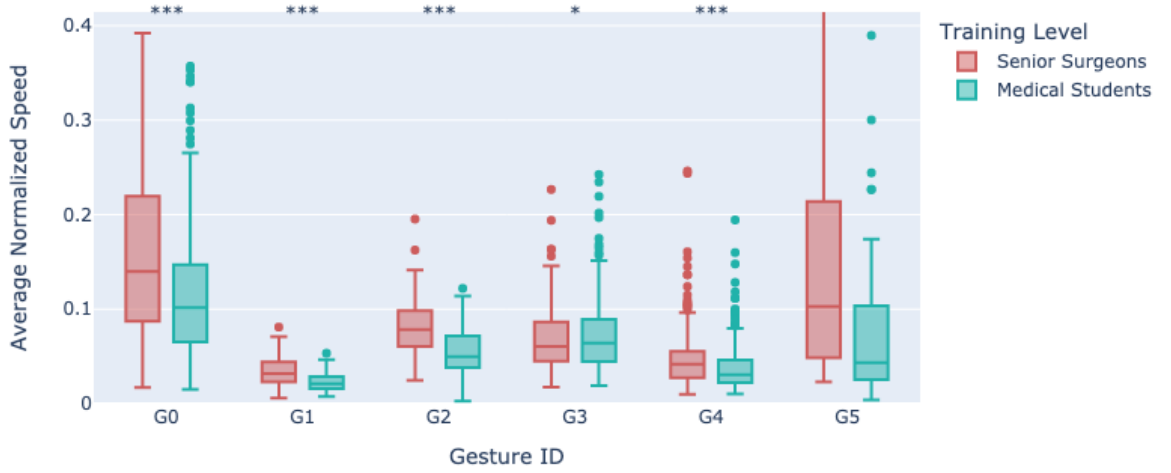
Figure 7.6: Box plots illustrating the normalized path lengths of the left and right hands across different surgical gestures (G0 to G5) during suture procedures. These procedures were performed on a tissue simulator by both attending surgeons and medical students.

Path Length: For each surgical gesture presented in Figure 7.6, we reported the normalized path lengths. The comparative analysis between attending surgeons and medical students across all surgical gestures reveals a nuanced performance difference. While attending surgeons demonstrate significantly shorter path lengths in certain gestures (G0, G1 for the left hand and G0, G3, G4 for the right hand), medical students show lower path length for gesture G5 with the right hand. For other gestures, the performance between the two groups is comparable.

Gesture Speed and Acceleration: Figures 7.7 and 7.8 depict the average normalized speed and average normalized acceleration, respectively. As expected, attending surgeons performed most surgical gestures (G1, G2, G3, G4, G5 for the left hand; G0, G1, G2, G4, G5 for the right hand) at a faster pace. Additionally, attending surgeons demonstrated higher acceleration compared to medical students for gestures G1, G2, G3, G4, and G5 with the left



(a) Left Hand

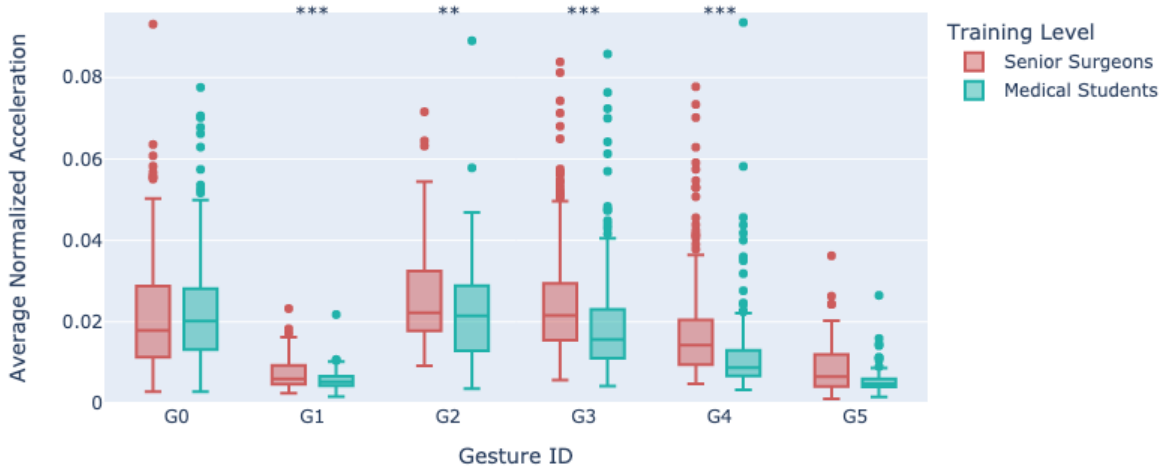


(b) Right Hand

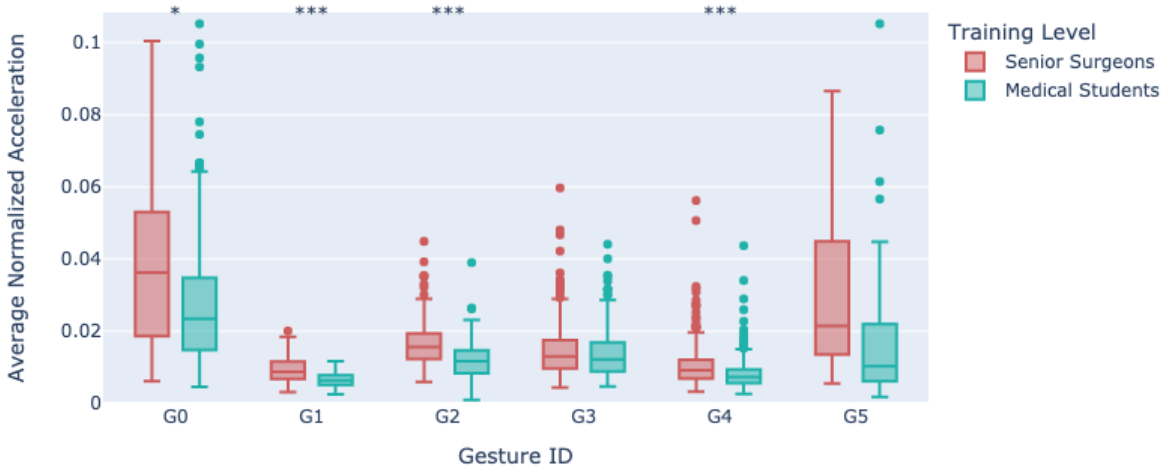
Figure 7.7: Box plots illustrating the averaged normalized speeds of the left and right hands across different surgical gestures (G0 to G5) during suture procedures. These procedures were performed on a tissue simulator by both attending surgeons and medical students.

hand, and for all surgical gestures with the right hand. These findings clearly indicate the attending surgeons' expertise, efficiency, and confidence in performing complex surgical gestures, likely developed through extensive practice and experience, resulting in refined and efficient movements.

Gesture Smoothness: The box plots in Figure 7.9 display the gesture smoothness performance metric. Attendingsurgeons exhibit significantly higher values for G1 (passing the needle through the material), G3 (performing an instrumental tie), and G4 (laying the knot) with the left hand, as well as for G1 (passing the needle through the material), G2 (pulling the suture), and G3 (performing an instrumental tie) with the right hand.



(a) Left Hand



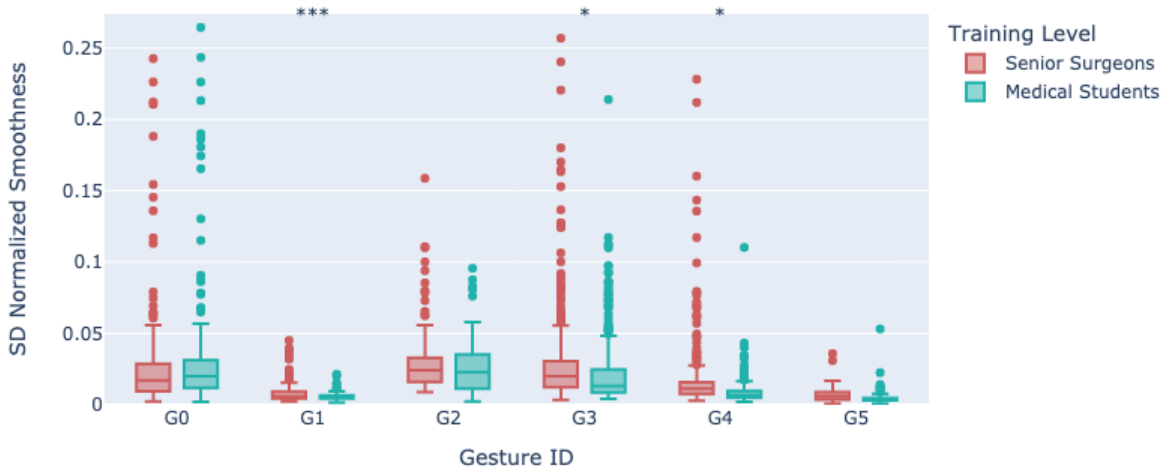
(b) Right Hand

Figure 7.8: Box plots illustrating the averaged normalized accelerations of the left and right hands across different surgical gestures (G0 to G5) during suture procedures. These procedures were performed on a tissue simulator by both attending surgeons and medical students.

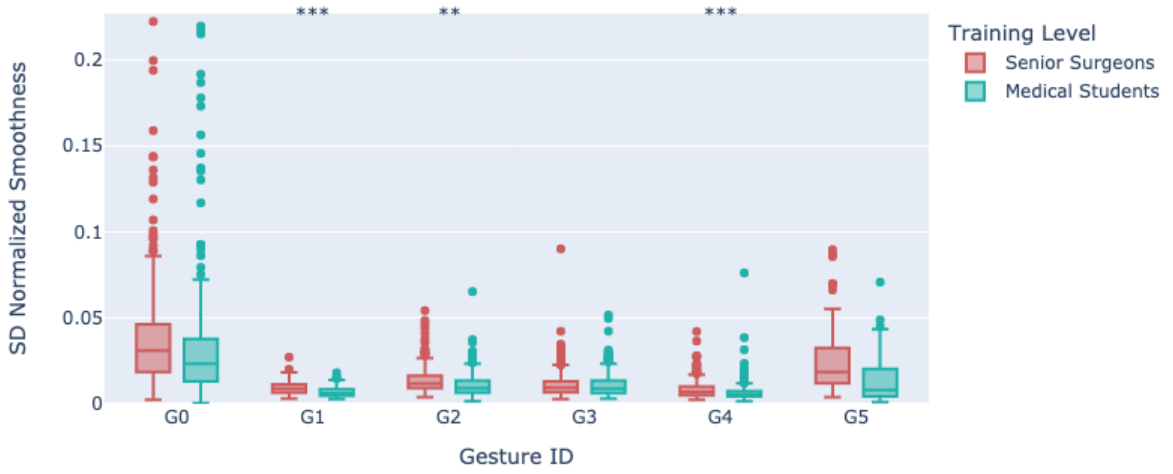
Gesture Curvature: We reported the average normalized curvature in Figure 7.10. No clear pattern was observed across the different surgical gestures for either hand. However, significantly higher mean values were noted for attending surgeons performing G0 with the left hand and G0 and G3 with the right hand. Conversely, medical students exhibited significantly higher mean values for G3 with the left hand and G1 with the right hand.

Summary and Implication for Surgical Training

The overall analysis reveals distinct performance on certain metrics between attending surgeons and medical students across multiple surgical gestures. Attending surgeons performed most surgical gestures in less time, at higher speeds and accelerations. They also



(a) Left Hand

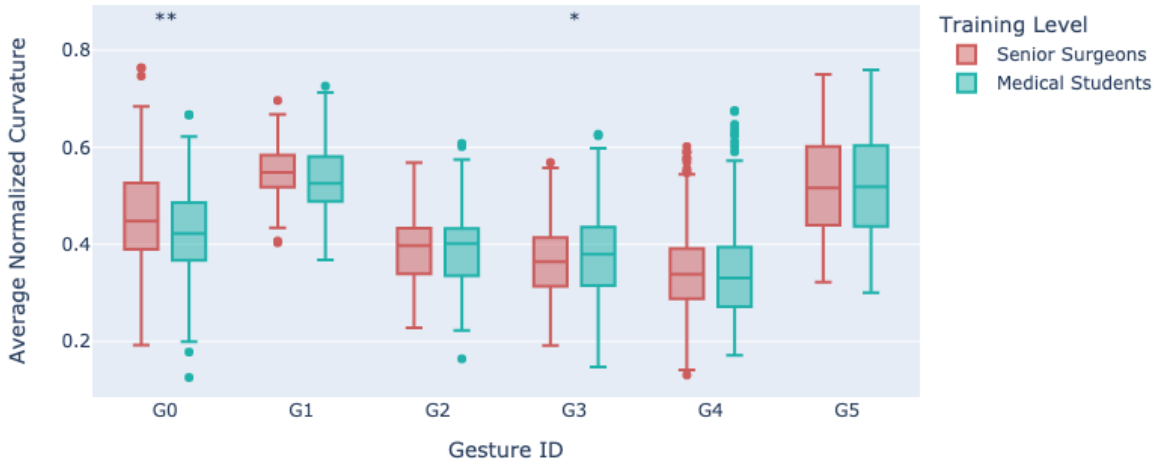


(b) Right Hand

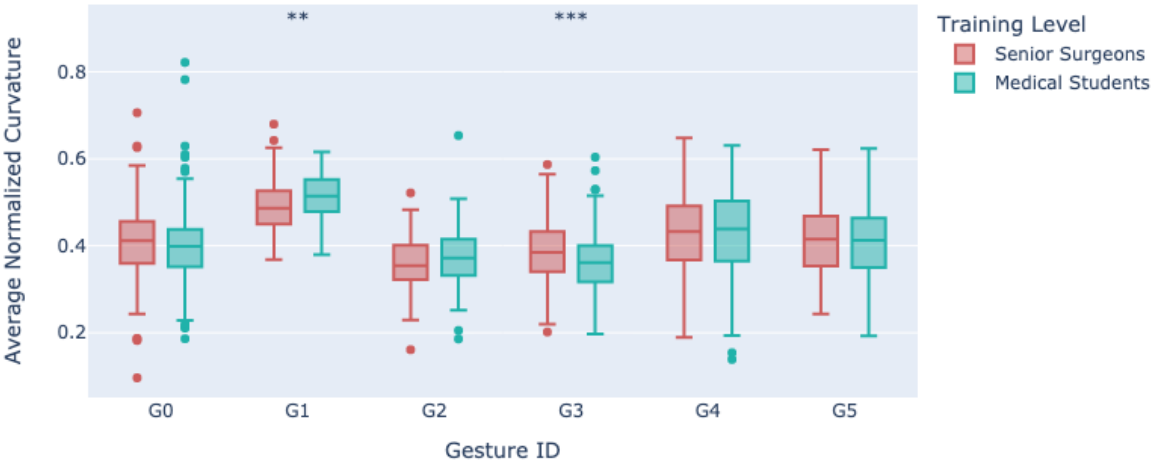
Figure 7.9: Box plots illustrating the standard deviation of the gesture smoothness performance metric for the left and right hands across various surgical gestures (G0 to G5) during suturing procedures. These procedures were conducted on a tissue simulator by both attending surgeons and medical students.

exhibited shorter path lengths and demonstrated higher gesture smoothness for certain surgical gestures. These findings offer valuable insights into the proficiency levels of experienced surgeons and highlight specific areas where medical students lag. These informations can directly inform surgical training by developing targeted tools that focus on improving these specific skills among medical students.

Automatic surgical gesture recognition, combined with performance metric analysis, can significantly enhance the development of training systems that monitor student performance in real time. As students engage in surgical tasks, the system can automatically identify and assess their gestures, comparing them to benchmarks set by experienced surgeons. For example, if the system identifies that a student's gestures are slower or less precise than those of an expert, it can immediately provide tailored feedback, offering specific adjustments or exercises to target those deficiencies. This real-time feedback loop is



(a) Left Hand



(b) Right Hand

Figure 7.10: Box plots illustrating the average normalized curvature for the left and right hands across various surgical gestures (G0 to G5) during suturing procedures. These procedures were conducted on a tissue simulator by both attending surgeons and medical students.

crucial for helping students rapidly correct errors, refine their techniques, and accelerate their learning process.

7.6 Discussions

7.6.1 Implications of Findings

The proposed MGRFormer framework demonstrates state-of-the-art performance in surgical gesture recognition, significantly outperforming existing methods on the VTS dataset. This achievement underscores the importance of leveraging multimodal data, specifically

the integration of kinematic and video modalities, to enhance surgical gesture recognition accuracy. By combining these modalities, the MGRFormer framework effectively captures complementary patterns from both data types, providing a more comprehensive understanding of surgical gestures. This capability is particularly valuable for developing advanced surgical training systems that can provide more accurate surgical gesture predictions.

Furthermore, effectively leveraging multimodal data is crucial for achieving optimal performance by making the most of each modality. A major contribution of this study is the introduction of a multimodal refinement module, which significantly enhances model accuracy. The clear performance gap between methods that use refinement and those that do not underscores the importance of this technique for accurate surgical gesture recognition.

The implications of this study extend beyond improved gesture recognition accuracy. By coupling performance metrics with surgical gesture recognition, the system can provide objective and targeted feedback, facilitating the identification of specific skill gaps for medical students. This capability will allow for the development of tailored training programs that can adapt to the individual needs of students.

Moreover, the MGRFormer's performance on the VTS dataset, which includes variability in environmental conditions such as lighting, occlusions, and minor camera displacements between recordings, demonstrate its adaptability to real-world surgical environments. This ability to perform well under varying conditions suggests that the model could be effectively adopted for real-world application.

7.6.2 Limitations

Despite the promising outcomes, several limitations need to be acknowledged that could impact the broader applicability of the MGRFormer framework.

A primary limitation is the computational complexity involved in training MGRFormer. The model's architecture, comprising two encoders and multiple decoders, demands substantial computational resources and time due to the large number of parameters that need to be optimized. Consequently, training such a model requires specialized hardware, which could limit its accessibility. Additionally, the model's complexity may hinder its deployment on wearable devices that have limited processing power and battery life.

Another significant limitation lies in the dependence on annotated surgical data for training the models. Acquiring high-quality annotated data is a time-consuming and expensive process, as it requires domain expertise to ensure the accuracy of the annotations.

A third limitation concerns the reliance on kinematic data for computing performance metrics to provide objective feedback to medical students on various aspects of their performance. In this study, kinematic data was collected using specialized and expensive equipment, such as sensor-based gloves that track hand movements. These devices, while providing motion data, are often cumbersome and uncomfortable for users, potentially hindering natural hand movements and interactions during surgical simulations. Additionally, the

high cost and maintenance requirements of such specialized equipment pose challenges for widespread adoption in educational settings, particularly for institutions with budget constraints.

Another significant challenge with the MGRFormer framework is the dependency on the simultaneous availability of both kinematic and video data for making predictions. In practice, ensuring that both types of data are consistently available can be difficult, particularly in real-world or clinical settings where data capture conditions are less controlled. The absence of either data type at any point can severely impair the model's ability to make accurate predictions, highlighting a critical vulnerability of the approach. This dependency can be a major barrier to broader implementation, especially in scenarios where kinematic data might be missing due to equipment failure, data loss, or limited access to specialized sensors. The need for both data types to be present and synchronized at all times complicates the deployment and reduces the robustness of the MGRFormer framework in environments with inconsistent data availability.

Lastly, the use of static view cameras in capturing video data for surgical gesture recognition. Static cameras, typically positioned at a fixed point in the simulation environment, limit the field of view and may fail to capture the intricate details of hand movements and instrument handling from multiple angles. This restricted perspective can result in occlusions, where critical gestures or tool manipulations are partially or fully obscured, leading to a loss of valuable visual information necessary for accurate gesture recognition. These limitations reduce the robustness and reliability of the recognition system in diverse or dynamic surgical settings where conditions frequently change. As a result, the current approach's dependency on static view recordings may not adequately capture the complexity and variability of real-world surgical procedures, limiting its effectiveness in providing comprehensive feedback in educational or clinical contexts.

7.6.3 Future Directions

To enhance the applicability and effectiveness of the MGRFormer framework, future research should focus on several key areas that address the identified limitations and expand upon the findings of this study.

Firstly, the MGRFormer model should be evaluated on additional surgical simulation datasets that encompass different surgical procedures to assess its effectiveness across diverse surgical contexts. This evaluation would require the collection of new datasets representing different surgical tasks, each featuring a large number of participant with varying surgical skills, and varying environment setting in order to develop generalizable models and effective models for real-world application.

To address the challenge of computational complexity, future work could explore more efficient architectures or model compression techniques, such as knowledge distillation, pruning, or quantization. These approaches could reduce the number of parameters and computational demands, leading to more efficient inference and deployment on resource-constrained devices.

To mitigate the reliance on annotated surgical data, future research should investigate more the application of unsupervised, or self-supervised learning approaches. These methods can leverage large amounts of unlabeled data for pre-training the model, which can subsequently be fine-tuned with a smaller annotated dataset. Such strategies would reduce the dependency on costly expert-annotated data, making the framework more scalable and feasible for widespread adoption. For instance, contrastive learning techniques could be employed to learn useful feature representations from unlabeled video data, enabling the model to better understand surgical gestures without extensive manual annotation.

Another promising direction involves the used of more advanced data collection methods, such as first-person view (FPV) perspectives. Current systems relying on static view cameras are limited by their fixed positions and potential occlusions, which can obscure critical gestures and tool manipulations. By incorporating FPV cameras mounted on surgical instruments or practitioners, the system could capture more detailed and dynamic views of hand movements and instrument handling. This would allow for a richer understanding of the surgical workflow, ultimately improving the system's robustness in recognizing and analyzing surgical actions. Furthermore, FPV setups are particularly beneficial in the context of surgical simulations, where replicating the complexity and variability of real-world procedures is crucial for effective training. FPV cameras can easily be integrated into simulation environments, capturing detailed and immersive perspectives that closely mimic the practitioner's view during surgery. This adaptability makes FPV ideal for advanced simulations involving complex, multi-step tasks and varied environments, thereby enhancing the realism and educational value of the simulation. Additionally, FPV can also be relevant for actual surgical operations, providing continuous, real-time perspectives that adapt to the surgeon's movements and capturing intricate details that static cameras might miss. This capability allows for better data collection even in the complex and variable conditions of live surgeries, supporting more effective training, evaluation, and performance analysis across both simulated and real-world contexts.

To further enhance the feedback provided to medical trainees, future research should also focus on developing more specialized performance metrics using advanced computer vision tools. Current metrics derived from kinematic data often require expensive and cumbersome equipment, which may not be practical in all settings. Instead, leveraging computer vision techniques to analyze video data could provide a less intrusive and more cost-effective way to measure key performance indicators. For instance, hand pose estimation and surgical tool trajectory analysis can be used to compute performance metrics. Specifically, hand pose estimation can facilitate the design of more advanced metrics, such as precise hand positioning, finger movement coordination, and joint angle variability during critical maneuvers, which are indicative of skill levels and proficiency. These advanced metrics will allow for a more detailed and nuanced assessment of a medical student's skills, capturing subtleties of performance without relying on specialized hardware.

7.7 Conclusion

This chapter introduces MGRFormer, a novel Transformer-based multimodal framework designed for the task of surgical gesture recognition. MGRFormer incorporates an innovative multimodal refinement module that effectively leverages the complementary information between kinematic and video data during the refinement stage. Extensive experiments on the VTS dataset demonstrate that MGRFormer outperforms by a large margin existing multimodal approaches and traditional fusion techniques, achieving state-of-the-art performance. Our results highlight superior recognition accuracy across various combinations of kinematic and video modalities, including frontal and side views, using ResNet-18 and I3D features.

Our findings underscore the critical role of integrating data from multiple sources to enhance surgical gesture recognition systems, providing a more comprehensive understanding of surgical actions. Moreover, we demonstrate the importance of effectively leveraging multimodal data to maximize the contribution of each modality. Additionally, we emphasize the significant impact of incorporating a refinement module in improving the performance of surgical gesture recognition systems.

We also present a comprehensive statistical analysis comparing the performance of attending surgeons and medical students across key metrics for all defined surgical gestures. This analysis reveals notable differences in performance on specific metrics for certain gestures, providing insights into varying levels of proficiency. Consequently, combining surgical gesture recognition systems with performance metrics calculated on the predicted gestures could facilitate the development of educational tools that provide granular, gesture-level feedback, ultimately enhancing surgical training and skill acquisition.

Chapter 8

Gesture Recognition in Surgical Simulation Training

Contents

This chapter introduces two novel datasets for surgical gesture recognition, addressing the limitations of existing datasets. The first dataset contains video recordings of attending surgeons and surgical residents performing the peg transfer task, while the second dataset features first-person video recordings of suturing tasks performed by both attending surgeons and medical students. For the peg transfer dataset, we conducted both unimodal and multimodal benchmarks. Additionally, we validated the MGRFormer framework, introduced in Chapter 7, within the multimodal benchmark setting. For the suturing dataset, we performed unimodal benchmark.

In Section 8.1, we discuss the importance of surgical simulation training and highlight the need for collecting new datasets to advance surgical gesture recognition in this context. Section 8.2 reviews the existing datasets and their limitations in the context of surgical gesture recognition. Subsequently, Section 8.3 introduces the two collected datasets. Next, Section 8.4 presents the experimental results for both datasets. Section 8.5 offers a comparative statistical analysis of surgical gestures performed by attending surgeons versus medical students across both datasets. Lastly, Section 8.6 discusses the implications, limitations, and future directions of our work, and Section 8.7 summarizes our contributions.

8.1 Introduction

Medical simulation has become an important part of modern medical education, providing a safe, controlled environment for medical students and professionals to develop and refine their skills without risk to patients. This is particularly important in surgical training for several reasons. First, surgical procedures are becoming increasingly complex, requiring more hands-on experience. Simulation sessions can help by allowing medical students to repeatedly perform complex tasks until they achieve the necessary level of dexterity and confidence. Furthermore, the traditional apprenticeship model of surgical training, often summarized as "see one, do one, teach one," is no longer adequate in today's medical landscape, where patient safety is paramount. Opportunities for trainees to practice and learn from mistakes on actual patients are limited. In this regard, simulation bridges the gap by providing a risk-free environment where errors can occur and be corrected without jeopardizing patient care.

In the preceding chapter, we introduced a novel multimodal deep learning approach for automatic surgical gesture recognition, which significantly outperformed the state-of-the-art on the VTS dataset [222]. This chapter serves as a continuation of that work, building upon the methodological advancements presented earlier. Here, we shift our focus to the data itself—specifically, the limitations of existing datasets and the need for more diverse and realistic data to further advance the field.

The incorporation of such advanced technologies into simulation-based training holds great promise for enhancing educational outcomes by offering objective performance metrics and instantaneous feedback. However, the success of these innovations is heavily dependent on the availability of comprehensive, high-quality datasets. Despite notable progress, the research community continues to face challenges due to the limited availability of diverse public datasets necessary for developing surgical gesture recognition meth-

ods.

In this chapter, we introduce two new datasets designed to address the limitations of existing datasets. Collected at the PRESAGE medical simulation center at the University of Lille, these datasets focus on two surgical tasks: peg transfer and suturing. The peg transfer task was performed by both attending surgeons and surgical residents multiple times. The dataset includes videos of the procedures and the corresponding surgical tool trajectories, tracked using a YOLOv8 model. The second dataset comprises first-person video recordings capturing attending surgeons and medical students performing a suturing procedure multiple times. We conducted unimodal and multimodal benchmarks for surgical gesture recognition on the peg transfer dataset, as well as a unimodal benchmark for surgical gesture recognition on the suturing dataset.

In both datasets, the ASFormer architecture [185] significantly outperformed state-of-the-art methods in the unimodal setting. Moreover, our previously introduced multimodal MGRFormer framework achieved superior performance on the peg transfer dataset for the multimodal setting.

Additionally, as done in the preceding chapter with the VTS dataset, we conducted a comparative statistical analysis on the defined surgical gestures across both datasets to evaluate performance differences between attending surgeons and medical students. The aim will be to combine surgical gesture predictions and the computation of performance metrics to improve surgical simulation training by providing objective feedback at the gesture level.

The key contributions of this chapter are summarized as follows:

1. We introduce two datasets focused on peg transfer and suturing tasks, collected at the PRESAGE medical simulation center. The peg transfer dataset includes both video recordings and tool trajectory data, while the suturing dataset features first-person video recordings.

2. We establish comprehensive unimodal and multimodal benchmarks for surgical gesture recognition on the peg transfer dataset, and a unimodal benchmark on the suturing dataset.

3. Our MGRFormer framework demonstrates state-of-the-art performance in the multimodal setting on the peg transfer dataset.

8.2 Related Work

Recent advancements in surgical gesture recognition have been significantly influenced by the introduction and evaluation of various datasets, each contributing to the field but also facing distinct challenges and limitations.

The JIGSAWS (JHU-ISI Gesture and Skill Assessment Working Set) [164] dataset is a widely used benchmark for surgical gesture recognition and surgical skill assessment. It

consists of kinematic and video data collected on three surgical simulation tasks (suturing, needle-passing, and knot-tying) performed on a da Vinci robotic surgical system. A key advantage of this dataset is its multimodal nature, allowing for the development of more robust and accurate models for surgical gesture recognition by leveraging complementary data sources. However, the dataset has several limitations. Firstly, the dataset includes only eight surgeons with varying skill levels (novice, intermediate, expert), significantly limiting the diversity and representativeness of the data. This small sample size raises concerns about the generalizability of models trained on this dataset, as these models may be prone to overfitting and might struggle when applied to data from other surgeons with different styles or skill levels. Furthermore, the fixed camera perspective in the JIGSAWS dataset restricts its ability to capture the full range of surgeon gestures and tool interactions—an especially important factor for complex tasks like suturing.

In [30], the authors evaluated their proposed framework using two public datasets containing both kinematic data and video recordings from a fixed camera perspective. These datasets, collected using the da Vinci Research Kit (dVRK) platforms at two different centers, consist of a peg transfer task performed by the same user on each platform, with 12 recorded sequences per site. The small size of the dataset as well as being limited to a single user pose significant challenges. The homogeneity of the dataset may not adequately capture the variability in surgical gestures across different users and environments, potentially leading to overfitting and limited generalizability of the model to other surgeons with varying techniques and styles. Moreover, the peg transfer exercise in this dataset is a simplified version of the task described in the Fundamentals of Laparoscopic Surgery (FLS) program [224]. By limiting the task to the transfer of a single peg from left to right, the dataset fails to capture the full complexity of the FLS peg transfer task, which involves the multiple transfers of six pegs. This simplification neglects key aspects of the task, such as the repetitive nature and the coordination challenges, which are crucial for assessing a surgeon's skill. As a result, models developed using these datasets may not be adequately tested on the broader range of skills required for laparoscopic surgery, potentially limiting their effectiveness and generalizability to more complex surgical tasks and real-world applications.

Similarly, Gazis et al. [225] introduced two datasets comprising training session videos on two fundamental laparoscopic tasks: peg transfer and knot tying. These tasks were performed 2-3 times by a group of 15 surgical trainees, resulting in a total of 40 trials for each task. The peg transfer task involved the placement of four cylindrical pegs onto a pegboard. The first two pegs were placed directly on the pegboard using either the left or right laparoscopic tool, while the remaining two pegs were first transferred between tools before being placed on the board. This version of the peg transfer task differs in some notable ways from the peg transfer exercise included in the FLS program. In the FLS exercise, the task typically involves six pegs rather than four. Moreover, the FLS exercise is structured to require that all six pegs be transferred from one side of the pegboard to the other, using both hands, and then transferred back to their original side. Another limitation is that the surgical procedures were exclusively performed by surgical trainees, without including attending surgeons. This restricts the ability to develop educational tools based on this dataset that could compare the performance of individual trainees to a benchmark set by experienced surgeons, which would provide more meaningful assessments and targeted

guidance, as we proposed to do in the previous chapter.

On the other hand, the VTS dataset [222], previously used in the preceding chapter to evaluate our proposed multimodal deep learning approach, offers several advantages. Notably, it includes a large group of participants: 12 attending surgeons, 11 medical students, and 1 surgical resident, each performing a suturing procedure twice on two different tissue simulators—tissue paper and a rubber balloon, resulting in a total of 96 surgical procedures. The balanced distribution between attending surgeons and medical students facilitates a comparative analysis of surgical gestures between the two groups across various performance metrics, contributing to the development of educational tools for medical students. Moreover, the dataset features both kinematics data and video data with two different views (close-up and side view) of the surgical procedures, enabling a richer and more comprehensive analysis. A key strength of this dataset, is that it includes a comprehensive view of the participants as they perform the suturing tasks. This allows for the observation of the entire performance, including the participant's hands, tools, and the complete surgical simulation environment, providing a more holistic perspective of the procedure. This enables a richer multimodal analysis, where the interaction between different aspects of the procedure can improve the accuracy of surgical gesture recognition systems. However, the dataset is not without limitations. Both cameras are fixed in position, leading to frequent occlusions during the procedures. This static camera setup poses challenges when attempting to replicate the dataset's conditions in more complex environments, such as an operating room, where dynamic and variable conditions are common.

These existing datasets have significantly advanced the field of surgical gesture recognition, each bringing valuable insights while also facing inherent limitations. To address some of these gaps, we introduce two new datasets on two surgical tasks: peg transfer and suturing. Unlike previous peg transfer datasets [30, 225], our dataset includes a larger and more diverse participant pool, featuring 11 attending surgeons and 14 surgical residents, each performing the task between 1 and 4 times, resulting in a total of 68 recorded procedures. Notably, this dataset exclusively contains video recordings of the exercise, in contrast to datasets like JIGSAWS, VTS, and those using the dVRK platform, which also incorporate kinematic data. These existing datasets rely on specialized equipment to capture kinematic data, such as the da Vinci robotic system or gloves with embedded sensors, which can impose significant constraints on scalability and generalizability.

Instead, we propose to extract the trajectory of the surgical tools directly from video recordings using a YOLOv8 deep learning model. This approach is more flexible and less intrusive, as it avoids the need for expensive robotic systems or wearable sensors, enabling a more natural and adaptable data collection process. Additionally, our peg transfer protocol aligns with the standards of the FLS program, involving multiple transfers of six pegs from left to right and vice versa.

The second dataset that we collected includes video recordings of attending surgeons and medical students performing a suturing task, captured from a first-person view (FPV). The inclusion of FPV footage addresses some of the critical limitations associated with fixed camera perspectives in existing datasets like JIGSAWS and VTS [164, 222]. Traditional fixed viewpoints often lead to occlusions, limited coverage of fine motor gestures, and a lack of detailed visualization of hand-tool interactions, all of which are vital for accurate

gesture recognition and assessment. In contrast, an FPV perspective offers an immersive view closely aligned with the surgeon's visual attention, capturing intricate tool movements and subtle hand coordination in a manner that better replicates real-world conditions.

8.3 Datasets

8.3.1 Peg Transfer Dataset

Peg Transfer Task Description

The peg transfer task is a fundamental exercise in the FLS program. It has been designed to evaluate a surgeon's dexterity, hand-eye coordination, and precision in using surgical instruments. During this exercise, participants are required to use two graspers to transfer six pegs from one side of a pegboard to the other and then back.

The protocol we followed for this dataset aligns with the FLS guidelines: (1) The pegs are initially placed on the left side of the pegboard. (2) The participant, using their non-dominant hand, lifts each peg from the left side and passes it to their dominant hand, which then places the peg on the right side. Once all six pegs have been transferred, the process is repeated in reverse, with the pegs being moved from the right side back to the left. (3) The task is considered complete when all pegs have returned to their original positions.

Dataset Overview

The dataset included 25 participants performing a peg transfer exercise between 1-4 times, resulting in a total of 68 procedures. The participants included, 11 attending surgeons and 14 surgical residents. The duration of each procedure ranges from 1 to 7 minutes. Each procedure was recorded with high-resolution video, capturing the field of view to ensure that every movement, gesture, and interaction with the pegs is visible for analysis.

The peg transfer task was subdivised into the following ten distinct gestures:

- G0: "the background gesture"
- G1: "reach for peg with the left grasper"
- G2: "lift the peg with the left grasper"
- G3: "transfer the peg from the left to the right grasper"
- G4: "place the peg into the pegboard with the right grasper"
- G5: "reach for peg with the right grasper"
- G6: "lift the peg with the right grasper"
- G7: "transfer the peg from the right to the left grasper"
- G8: "place the peg into the pegboard with the left grasper"
- G9: "peg drops"

Figure 8.1 presents a sequence of images illustrating the first five surgical gestures in-

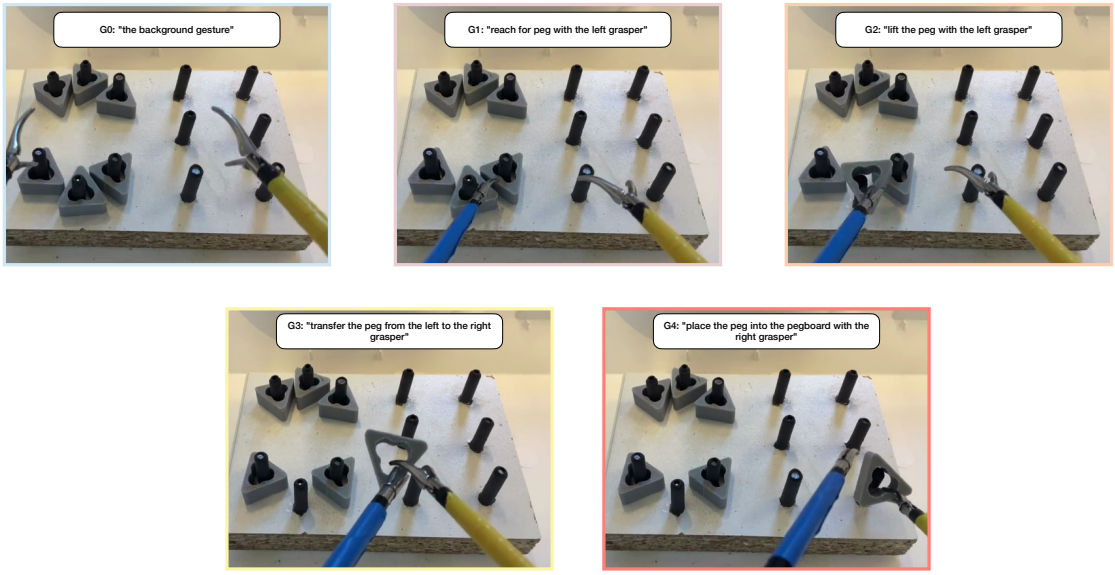


Figure 8.1: Illustration of the first five surgical gestures during the peg transfer task, focusing on the movement of pegs from the left to the right side. The depicted gestures include: (G0) "the background gesture", (G1) "reaching for the peg with the left grasper", (G2) "lifting the peg with the left grasper", (G3) "transferring the peg from the left to the right grasper", and (G4) "placing the peg into the pegboard with the right grasper". The images are ordered progressively to illustrate the procedural flow.

volved in transferring a peg from the left to the right side of the pegboard. These gestures capture the fundamental actions required to complete half the peg transfer task, segmenting the entire procedure into discrete, meaningful actions. This granularity is important for accurate gesture recognition and analysis, as each gesture reflects specific skills such as precision, coordination, and control, all of which are important in surgical contexts. The inclusion of G0 (the background gesture) and G9 (peg drops) is particularly important. G0 helps distinguish between meaningful surgical gestures and periods of inactivity or transitions, reducing noise and improving the clarity of analysis. G9, representing peg drops, highlights critical errors during the task, offering insight into skill levels and error-handling strategies. Together, these gestures provide a comprehensive representation of the peg transfer exercise, aligning with existing protocols in surgical training and making the dataset directly comparable to other studies and resources.

Surgical Tool Motion Extraction

Motivation: In addition to the recorded videos of the peg transfer procedure, we propose to extract surgical tool motion data, which offers several advantages for surgical gesture recognition.

Motion data from surgical tools directly reflects the surgeon's actions, enabling more precise and accurate recognition of surgical gestures. Furthermore, performance metrics derived from this data, such as path length, velocity, smoothness, allow for objective com-

parisons between attending surgeons and residents on the recognized gestures. This data facilitates more detailed assessments of surgical skill.

Compared to video data, using motion data for gesture recognition offers several key benefits. Firstly, motion data is inherently lower-dimensional and involves smaller data volumes, capturing essential parameters like the position and movement of surgical tools. On the other hand, video data is high-dimensional, consisting of large volumes of pixel information across frames, often containing irrelevant or redundant content. Motion data is more focused and directly correlated with surgical gestures, reducing the need for complex feature extraction and requiring less computational power. Additionally, unlike video data, motion data is less susceptible to noise from external factors such as occlusions, lighting variations, or camera positioning. This makes it a more reliable source for surgical gesture recognition, contributing to more consistent and accurate model performance.

The combination of both tool motion and video data provides complementary information that can enhance the robustness of surgical gesture recognition models. While motion data offers direct insights into tool trajectories, video data captures contextual information, such as the overall surgical environment and interactions between tools and tissue. Integrating both sources of information can lead to better generalization, improved recognition accuracy, and the ability to handle challenging scenarios like partial occlusions or ambiguous hand movements that might not be fully resolved by either modality alone. In this study, we aim to test whether this fusion of modalities leads to better performance than using each modality separately for surgical gesture recognition, building on our findings in the preceding chapter, where combining kinematic and video data resulted in significant performance improvements.

YOLOv8 Model: To achieve accurate tool motion extraction, we employ the YOLOv8 [226] model in conjunction with the ByteTrack tracking algorithm. Released by Ultralytics in January 2023, the company behind YOLOv5 [227], YOLOv8 offers significant improvements across multiple vision tasks, including object detection, segmentation, pose estimation, tracking, and classification. YOLOv8 is available in five scaled versions: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), and YOLOv8x (extra-large), allowing for tailored performance depending on computational resources and task complexity. For our study, we employed the YOLOv8m configuration.

YOLO (You Only Look Once) is a well-established family of object detection models known for its real-time performance and high detection accuracy. It has first been introduced in [228]. Unlike traditional object detection models that involve multiple stages (like region proposal, classification, and refinement), YOLO is a single-stage detector that directly predicts bounding boxes and class probabilities from the input image in one go. This architecture allows it to be extremely fast and efficient, making it suitable for real-time applications like surgical tool tracking, where high processing speed is essential.

We selected YOLOv8 specifically due to its state-of-the-art performance in both speed and accuracy. YOLOv8 introduces further optimizations such as dynamic anchor boxes, adaptive training optimizations, and better handling of edge cases where objects may be partially occluded or overlap. In our context, this is crucial as surgical tools often interact

closely, overlap, or move rapidly within the camera's field of view. YOLOv8's ability to maintain detection accuracy under such conditions makes it highly effective for identifying surgical tools across the frames in our video data.

However, detecting tools in each frame is just one part of the problem. To extract meaningful motion data over time, we need to ensure the detected tools are consistently tracked across frames. This is where the ByteTrack algorithm comes into play. ByteTrack is a robust multi-object tracking algorithm that excels in maintaining consistent identities for objects even in challenging scenarios involving occlusions, fast movements, and varying scales. It works by integrating both high-confidence and low-confidence detections, effectively bridging gaps where an object might be momentarily lost due to partial occlusion or rapid movement. This attribute is particularly valuable in surgical contexts, where the tracking of tools needs to be uninterrupted despite dynamic interactions and occlusions.

By leveraging YOLOv8 for detection and ByteTrack for tracking, we can reliably generate smooth and consistent tool trajectories, enabling a detailed analysis of surgical gestures. The resulting motion data provides critical insights into the movement dynamics of the tools, including metrics like position, speed, and tool trajectories, which directly contribute to recognizing and evaluating surgical gestures.

YOLOv8 Training Settings: For model development, we randomly sampled 985 images from video recordings of all procedures, capturing a diverse range of tool positions, orientations, and interactions with the environment. This approach ensures the model can generalize effectively across various scenarios. The images were annotated with bounding boxes using the CVAT tool, labeling the positions of the graspers. Of the annotated images, 800 were used for training, while 185 were reserved for testing. The model was trained for 1000 epochs with a batch size of 16, employing an early stopping criterion with a patience of 100 epochs.

YOLOv8 Model Performance: We evaluated the performance of the YOLOv8 model on several key metrics commonly used in object detection tasks.

Main Evaluation Metrics:

- **Precision**: Precision measures the proportion of correctly predicted bounding boxes among all predicted boxes. High precision indicates that the model produces fewer false positives.
- **Recall**: Recall measures the proportion of correctly predicted bounding boxes among all actual boxes. High recall means that the model captures most of the true objects.
- **mAP (mean Average Precision)**: mAP is a widely used metric in object detection that averages precision across multiple IoU (Intersection over Union) thresholds. It is a strong indicator of the model's overall detection accuracy.

The following table summarizes the performance of the YOLOv8 model on the test set:

Metric	Value
Precision	98.44%
Recall	96.54%
mAP@0.5 (IoU = 0.5)	98.66%
mAP@0.5:0.95 (Average)	69.03%

Table 8.1: YOLOv8 Performance Metrics on the Test Set

Discussion of Results: The YOLOv8 model demonstrated excellant performance across several key metrics. With a precision of 98.44%, the model exhibits a very low rate of false positives, indicating that nearly all detected bounding boxes correspond to actual surgical graspers. This high precision is particularly crucial in surgical environments where accuracy is paramount.

The recall rate of 96.54% suggests that the model is highly effective at capturing the vast majority of surgical graspers present in the test set. This high recall ensures that very few tools are missed, even in challenging conditions like occlusions or complex tool interactions.

The mAP@0.5 of 98.66% confirms the model's strong detection accuracy at an IoU threshold of 0.5, indicating that the bounding boxes predicted by the model closely match the ground truth. However, the mAP@0.5:0.95, which averages precision across more IoU thresholds, is 69.03%. This lower value suggests that while the model performs exceptionally well at a moderate IoU threshold, its accuracy diminishes when required to predict bounding boxes that must more precisely match the ground truth. This decrease in performance at higher IoU thresholds may be attributed to the challenges posed by the variability in tool sizes, shapes, and overlapping scenarios common in surgical environments.

Despite this, the model's overall performance, as indicated by its precision, recall, and mAP@0.5 scores, suggests it is highly suitable for the task of surgical tool tracking and motion extraction. The trade-off seen in the mAP@0.5:0.95 metric highlights areas for potential future refinement, such as improving the model's ability to handle more precise localization in challenging conditions.

These results underscore the YOLOv8 model's capability to deliver reliable and accurate tool tracking in real-time, reinforcing its role as a critical component of our surgical gesture recognition pipeline.

8.3.2 FPV Suturing Dataset

Dataset Overview

The suturing task is a fundamental component of medical training, particularly for medical students, surgical residents, and other healthcare professionals who need to develop essential skills in wound closure and surgical procedures.

Figure 8.2: Pupils Invisible glasses equipped with a camera for first-person video recording.

The dataset includes five participants: two attending surgeons and three medical students. Each participant performed the suturing task three times, resulting in a total of 15 procedures. The duration of each procedure ranges from 2 to 4 minutes. The task involved placing three interrupted sutures on a suture skin model with support, composed of silicone and foam. For this, each participant was equipped with three tools: a needle driver, surgical forceps, and suture scissors. Each procedure was recorded from a first-person perspective using Pupil Invisible glasses, a specialized eyewear designed for capturing first-person video footage. An illustration of the glasses is provided in Figure 8.2.

We propose to subdivide the suturing task into the following five distinct gestures:

- G0: "the background gesture"
- G1: "pass the needle through the material"
- G2: "pull the suture"
- G3: "perform an instrumental tie"
- G4: "cut the suture"

Figure 8.3 presents a sequence of five images, each depicting a distinct surgical gesture involved in the suturing task. We follow the same labeling protocol as introduced in the VTS dataset [222], with the exception that we chose to merge the gestures "perform an instrumental tie" and "lay the knot" from the VTS dataset into a single gesture, "perform an instrumental tie." This decision is justified by the fact that these two actions are closely related and often performed as part of a continuous sequence within the suturing process. The distinction between "perform an instrumental tie" and "lay the knot" can be ambiguous, leading to potential inconsistencies in annotations, which can negatively impact model performance. By consolidating these actions into a single category, we simplify the labeling process, reduce annotation noise, and maintain the key information needed for accurate gesture recognition.

Both datasets were labeled throughout the entire duration of each video procedure using the Computer Vision Annotation Tool (CVAT). Additionally, the protocol for both datasets received approval from the Institutional Review Board of the University of Lille, under reference number 2022-626-S108.

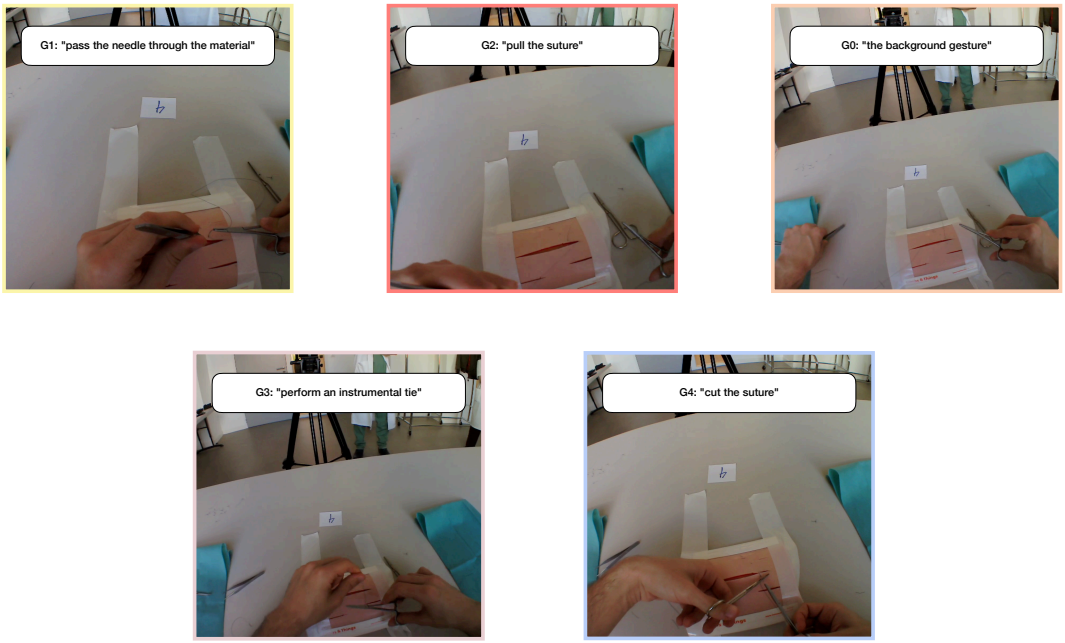


Figure 8.3: Illustration of the five surgical gestures during the suturing task. The depicted gestures include: (G0) "the background gesture", (G1) "pass the needle through the material", (G2) "pull the suture", (G3) "perform an instrumental tie", and (G4) "cut the suture". The images are ordered progressively to illustrate the procedural flow.

8.4 Surgical Gesture Recognition

8.4.1 Evaluation Metrics

As in the previous chapter, we will evaluate the different models trained on the two datasets for surgical gesture recognition using frame-wise and segmentation metrics. These metrics include accuracy, macro F1-score, Edit score, and segmental F1-score at thresholds of 10, 25, and 50.

8.4.2 Evaluation Framework

For the Peg Transfer and FPV Suturing datasets, we employed a subject-independent 5-fold and 3-fold cross-validation strategy, respectively, to evaluate the trained methods. In each fold, the datasets were divided into training and testing sets, ensuring that all data from each participant were exclusively included in either the training set or the testing set for that fold. We calculated and reported the mean value of each evaluation metric across all folds.

Method	Modality	Features	Acc	F1-Macro	Edit	F1@{10,25,50}		
LSTM [181]	Tools	✗	79.73	70.92	80.39	83.52	80.20	68.58
GRU [229]	Tools	✗	81.35	72.72	80.47	83.94	81.60	70.56
MS-TCN++ [200]	Tools	✗	63.84	54.24	62.14	67.00	61.71	45.65
ASFormer [185]	Tools	✗	**81.51**	**73.67**	**85.73**	**88.25**	**85.32**	**73.22**
LSTM [181]	Video	ResNet-18	71.00	62.39	68.11	72.51	68.69	57.82
GRU [229]	Video	ResNet-18	74.24	66.43	69.07	74.63	71.38	61.08
MS-TCN++ [200]	Video	ResNet-18	66.28	56.47	59.15	63.51	58.61	43.80
ASFormer [185]	Video	ResNet-18	**80.71**	**73.38**	**84.47**	**86.95**	**84.33**	**73.27**
LSTM [181]	Video	I3D	85.24	78.39	76.08	82.79	80.42	73.28
GRU [229]	Video	I3D	86.73	79.28	78.22	84.04	82.15	74.87
MS-TCN++ [200]	Video	I3D	80.39	72.52	74.45	80.13	77.63	68.53
ASFormer [185]	Video	I3D	**87.60**	**81.77**	**88.70**	**91.48**	**90.16**	**82.32**

Table 8.2: Unimodal surgical gesture recognition on the Peg Transfer dataset.

8.4.3 Peg Transfer

We carried out both unimodal and multimodal benchmarks using surgical tools trajectory and video data. For the video modality, we employed ResNet-18 and I3D extracted features, as proposed in the preceding chapter.

Unimodal

Table 8.2 presents the performance of four deep learning models: ASFormer, MS-TCN++, LSTM, and GRU across two modalities: surgical tools trajectory (Tools) and video (using ResNet-18 and I3D features). ASFormer consistently outperforms the other models across all evaluation metrics and modalities, demonstrating its robustness and adaptability for the task of surgical gesture recognition.

For the Tools modality, ASFormer surpassed GRU, the second-best performing model, by at least 0.95%, 5.26%, and 4.31% in macro F1-score, Edit score, and F1@10, respectively. There is a noticeable performance improvement when using I3D features compared to ResNet-18 features for all models in the case of the video modality, suggesting that the richer temporal features captured by I3D significantly enhance the model's effectiveness. For instance, ASFormer's F1-Macro score increases from 73.38% with ResNet-18 features to 81.77% with I3D features.

The performance of MS-TCN++ across all features and modalities is consistently lower compared to other models. For the Tools modality, MS-TCN++ achieves the lowest scores across all metrics, with an accuracy of 63.84%, an F1-Macro of 54.24%, and an Edit score

Figure 8.4: Color-coded illustration of surgical gesture recognition on the Peg Transfer dataset, comparing ground truth with MGRFormer$_{k \to v}$ predictions, trained using surgical tools trajectory and I3D features.

of 62.14%, indicating a difficulty to capture the temporal dynamics effectively compared to other models. When evaluated on video data with ResNet-18 features, MS-TCN++ continues to underperform, with an F1-Macro of 56.47% and an Edit score of 59.15%, further demonstrating its limitations in capturing temporal consistency for surgical gesture recognition.

While the performance of MS-TCN++ improves when employing I3D features, achieving better results than when using either the Tools modality or ResNet-18 features, it still lags significantly behind the other models, particularly ASFormer. This underscores that despite some improvement with richer temporal features, MS-TCN++ remains less effective compared to other models for surgical gesture recognition on the Peg Transfer dataset.

Lastly, it is important to note that although ASFormer performs comparably to GRU and LSTM in terms of frame-wise metrics like accuracy and F1-Macro, it significantly outperforms these models in segmentation metrics, which are critical for assessing temporal consistency and the model's ability to accurately segment sequences. For instance, in the Tools modality, while ASFormer and GRU have relatively close accuracy (81.51% vs. 81.35%) and F1-Macro (73.67% vs. 72.72%), ASFormer shows a marked improvement in the Edit score (85.73% vs. 80.47%) and segmentation F1 scores, such as F1@10 (88.25% vs. 83.94%). This pattern is consistent across other feature sets, including ResNet-18 and I3D, demonstrating ASFormer's superior capability in modeling temporal dynamics and segmenting surgical gestures more accurately.

Multimodal

In the multimodal setting, we evaluated the performance of the previously introduced MGRFormer across all combinations of single and double refinement regarding the fusion of the Tools modality with ResNet-18 and I3D features. Additionally, we reported the performance of ASFormer under multimodal fusion, specifically ASFormer (early) and ASFormer (late) for early and late fusion strategies, as well as MS-TCN++ (early) and MS-TCN++ (late).

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
MGR-Net [30]	72.87	62.97	66.10	71.21	66.46	53.53
MS-TCN++ (early)	71.58	60.89	60.83	66.29	62.94	49.67
MS-TCN++ (late)	69.31	57.80	62.94	66.81	63.10	49.13
ASFormer (early)	83.78	76.77	85.29	88.69	86.62	76.75
ASFormer (late)	83.43	76.31	78.79	84.17	81.98	72.46
MGRFormer $v \rightarrow k$	83.08	76.47	85.77	88.35	86.18	76.31
MGRFormer $k \rightarrow v$	**85.21**	77.92	86.87	89.55	88.06	**79.54**
MGRFormer $v \rightarrow v + k$	82.21	75.66	84.30	87.45	84.91	74.90
MGRFormer $k \rightarrow k + v$	84.39	77.13	83.28	87.15	85.24	77.08
MGRFormer $v \rightarrow k + v$	82.97	75.44	85.19	87.75	85.64	75.85
MGRFormer $k \rightarrow v + k$	85.03	**78.40**	**86.89**	**89.75**	**88.12**	79.50

Table 8.3: Multimodal surgical gesture recognition on the Peg Transfer dataset: Tools + ResNet-18.

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
MGR-Net [30]	72.57	63.19	66.24	71.44	67.10	54.49
MS-TCN++ (early)	82.79	74.15	78.20	83.10	80.73	72.51
MS-TCN++ (late)	83.55	75.54	77.09	82.81	80.95	72.64
ASFormer (early)	**88.73**	**83.08**	88.52	91.35	90.07	83.05
ASFormer (late)	87.46	81.22	80.59	86.25	84.54	77.19
MGRFormer $v \rightarrow k$	88.25	82.20	88.58	91.24	89.66	83.57
MGRFormer $k \rightarrow v$	88.57	82.40	88.66	91.62	90.29	82.97
MGRFormer $v \rightarrow v + k$	88.24	82.06	88.95	91.78	90.44	83.37
MGRFormer $k \rightarrow k + v$	88.05	82.00	84.66	88.95	87.33	81.07
MGRFormer $v \rightarrow k + v$	88.18	82.17	88.47	91.55	89.87	82.97
MGRFormer $k \rightarrow v + k$	88.32	82.18	**89.31**	**92.03**	**90.70**	**83.56**

Table 8.4: Multimodal surgical gesture recognition on the Peg Transfer dataset: Tools + I3D.

Figure 8.5: Color-coded illustration of surgical gesture recognition on the FPV Suturing dataset, comparing ground truth with ASFormer predictions, trained using I3D features.

These were compared alongside MGR-Net [30], a state-of-the-art multimodal deep learning model for surgical gesture recognition. All performances are reported in Tables 8.3 and 8.4.

For the combination of Tools and video with ResNet-18 features, MGRFormer significantly outperforms each modality when trained individually with the ASFormer model. Specifically, MGRFormer$_{k\rightarrow v+k}$ outperforms ASFormer trained on the Tools modality by 4.73%, 1.16%, and 1.50%, and ASFormer trained on ResNet-18 features by 5.02%, 2.42%, and 2.80%, in terms of macro F1-score, Edit score, and F1@10, respectively. When compared to MGR-Net, MGRFormer demonstrates a substantial performance gain, specifically outperforming MGR-Net by 15.43%, 20.79%, and 18.54% in macro F1-score, Edit score, and F1@10, respectively. Additionally, MGRFormer surpasses both the early and late fusion versions of ASFormer, highlighting the effectiveness of our proposed multimodal refinement module over traditional fusion techniques. It also achieves a significant performance margin over both early and late fusion versions of MS-TCN++, further validating its superior multimodal learning capabilities.

Among the various configurations of MGRFormer, MGRFormer$_{k\rightarrow v}$ exhibited the highest accuracy and F1@50, while MGRFormer$_{k\rightarrow v+k}$ achieved the best performance across the remaining metrics, solidifying its position as the most robust configuration for multimodal surgical gesture recognition.

For the fusion of surgical tool trajectories with I3D features, MGRFormer$_{k\rightarrow v+k}$ also showed superior performance compared to ASFormer models trained on individual modalities. Specifically, MGRFormer$_{k\rightarrow v+k}$ outperformed ASFormer trained on the Tools modality by 8.51%, 3.58%, and 3.78%, and ASFormer trained on I3D features by 0.41%, 0.61%, and 0.55%, in macro F1-score, Edit score, and F1@10, respectively. Furthermore, MGRFormer consistently outperformed MGR-Net again across all evaluation metrics. However, in this fusion setting, ASFormer (early) marginally outperformed MGRFormer$_{k\rightarrow v}$ by 0.16%, 0.68% in terms of accuracy and macro F1-score (frame-wise metrics). Nonetheless, for the remaining evaluation metrics (segmentation metrics), MGRFormer$_{k\rightarrow v+k}$ outperformed ASFormer (early) by 0.79%, 0.68%, 0.63%, and 0.51% in terms of Edit score F1@10, F1@25, and F1@50, respectively. Once again, MGRFormer maintained a significant performance margin over both early and late multimodal fusion versions of MS-TCN++.

Among the MGRFormer configurations, MGRFormer$_{k\rightarrow v}$ exhibited the highest accuracy and macro F1-score, with a slight lead over MGRFormer$_{k\rightarrow v+k}$, which achieved the best performance across the remaining segmentation metrics, solidifying its position as the most robust configuration in terms of segmentation metrics for multimodal surgical gesture

Method	Modality	Acc	F1-Macro	Edit	F1@{10,25,50}		
LSTM [181]	ResNet-18	67.48	48.81	41.58	46.05	41.30	30.47
GRU [229]	ResNet-18	67.35	49.15	30.94	32.18	28.07	19.07
MS-TCN++ [200]	ResNet-18	71.53	52.65	40.73	45.41	41.06	31.95
ASFormer [185]	ResNet-18	**78.96**	**65.71**	**72.57**	**72.52**	**68.14**	**59.04**
LSTM [181]	I3D	74.18	55.27	41.49	45.79	41.96	30.98
GRU [229]	I3D	74.56	57.19	42.99	46.04	41.81	30.63
MS-TCN++ [200]	I3D	77.48	60.21	48.77	52.85	46.35	33.27
ASFormer [185]	I3D	**82.41**	**69.61**	**82.14**	**81.99**	**73.71**	**61.11**

Table 8.5: Unimodal surgical gesture recognition on the FPV Suturing dataset.

recognition.

8.4.4 FPV Suturing

We evaluated the performance of ASFormer, MS-TCN++, LSTM, and GRU models using ResNet-18 and I3D features. The ASFormer consistently outperformed other models across all evaluation metrics and feature types, demonstrating its superior ability in surgical gesture recognition within the FPV setting. Notably, we can observed that all models performed better when employing I3D features, highlighting I3D's enhanced capability to capture spatiotemporal information.

Using ResNet-18 features, ASFormer achieved the highest scores across all metrics, with an accuracy of 78.96%, a macro F1-score of 65.71%, and an an Edit score of 72.57%. ASFormer also performed best across all overlap thresholds for F1 scores. MS-TCN++, the next best-performing model, showed considerably lower results, with an accuracy of 71.53%, F1-Macro of 52.65%, and an Edit score of 40.73%.

With I3D features, ASFormer achieved an accuracy of 82.41%, a macro F1-score of 69.61%, an Edit score of 82.14%, and an F1@10 score of 81.99%, significantly outperforming MS-TCN++, which, despite improved results with I3D features compared to ResNet-18, still lagged behind ASFormer with an accuracy of 77.48%, macro F1-score of 60.21%, and Edit score of 48.77%.

The significant difference in segmentation metrics between the ASFormer and other models across both type of features, highlights ASFormer's superior segmentation and refinement abilities in temporal boundaries, which are critical for precise gesture recognition.

Comparatively, results for the FPV Suturing dataset are generally lower than those for the Peg Transfer dataset, which is not unexpected given the challenges associated with the FPV suturing task. It is important to note that the FPV Suturing dataset features a smaller number of subjects (5 compared to 25 for the Peg Transfer dataset) and involves a first-

person view of a suturing procedure, which presents additional complexity. Despite these challenges, the results remain promising, highlighting the robustness of the ASFormer even under more difficult conditions. This underscores the encouraging potential of the model to handle complex and realistic surgical tasks, even if the performance metrics do not reach the levels observed in less challenging scenarios like the Peg Transfer dataset.

Figures 8.4 and 8.5 show the visual comparisons of the predictions generated by our proposed MGRFormer$_{k\to v}$ framework and the ASFormer architecture on the Peg Transfer and FPV Suturing datasets, respectively, compared to their respective ground truths. The MGRFormer$_{k\to v}$ integrates surgical tools trajectory data with I3D features, while ASFormer employs only I3D features. These visualizations highlight the temporal segmentation consistency of surgical gesture predictions for both architectures.

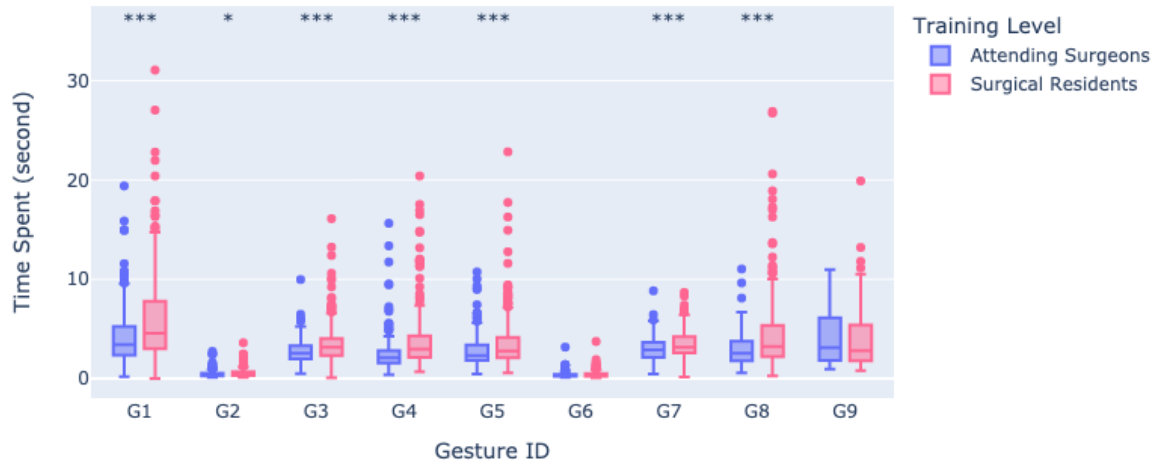


Figure 8.6: Box plots comparing the time spent (in seconds) to complete the different surgical gestures (G1 to G9) during peg transfer procedures by attending surgeons and surgical residents. Significance levels are indicated as follows: * p-value < 0.05, ** p-value < 0.01, and *** p-value < 0.001.

8.5 Surgical Gesture Analysis

In this section, we will conduct a comparative analysis of the performance between attending surgeons and medical students across the defined surgical gestures for the two introduced datasets. This approach follows the methodology applied to the VTS dataset in the preceding chapter for surgical gesture analysis.

For the Peg Transfer dataset, we report the computation of the gesture duration and frequency performance metrics for both attending surgeons and surgical residents across all surgical gestures. Additionally, we compute path length, gesture speed, acceleration, smoothness, and curvature based on the trajectories of both the left and right graspers. Note that we did not include performance metrics for the background gesture (G0), as this gesture typically occurs at the beginning and end of the peg transfer tasks and involves minimal movement of the graspers.

Figure 8.7: Box plots showing the frequencies of surgical gestures (G1 to G9) during peg transfer procedures, performed by both attending surgeons and surgical residents.

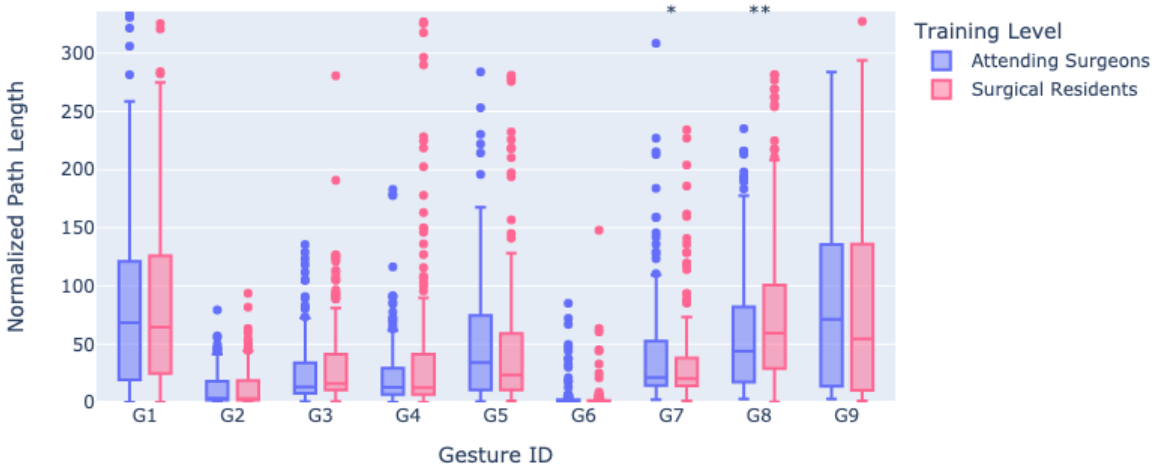
In the case of the FPV Suturing dataset, we report only the gesture duration and frequency metrics across all defined gestures due to the lack of data required to compute the additional performance metrics for both attending surgeons and medical students.

8.5.1 Peg Transfer

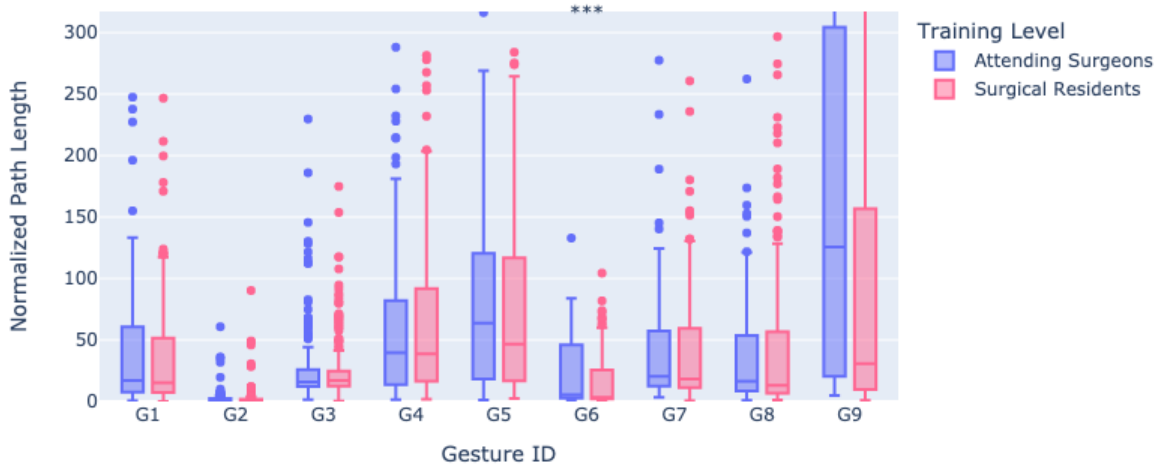
Gesture Completion Time: Figure 8.6 presents box plots comparing the time (in seconds) spent by attending surgeons and surgical residents to complete different surgical gesture (G1 to G9) during peg transfer tasks. Attending surgeons consistently demonstrated significantly shorter completion times compared to surgical residents across most surgical gestures. Specifically, significant differences were observed for gestures G1, G2, G3, G4, G5, G7, and G8.

Gesture Frequency: We reported in Figure 8.7, box plots showing the frequency associated with each surgical gesture. The plots reveal that surgical residents exhibit a significantly higher frequency for the gestures of transferring the peg from the left to the right grasper (G3) and placing the peg into the pegboard with the left grasper (G8).

Path Length: In Figure 8.8, the box plots illustrate the normalized path length across all surgical gestures. The comparative analysis between the two practitioner types reveals nuanced performance differences. Attending surgeons exhibited significantly shorter path lengths for gesture G8 in the left grasper's trajectory. In contrast, surgical residents demonstrated significantly shorter path lengths for gesture G7 in the left grasper's trajectory and G6 in the right grasper's trajectory. For other gestures, the performances between the two groups were comparable.



(a) Left grasper's trajectory

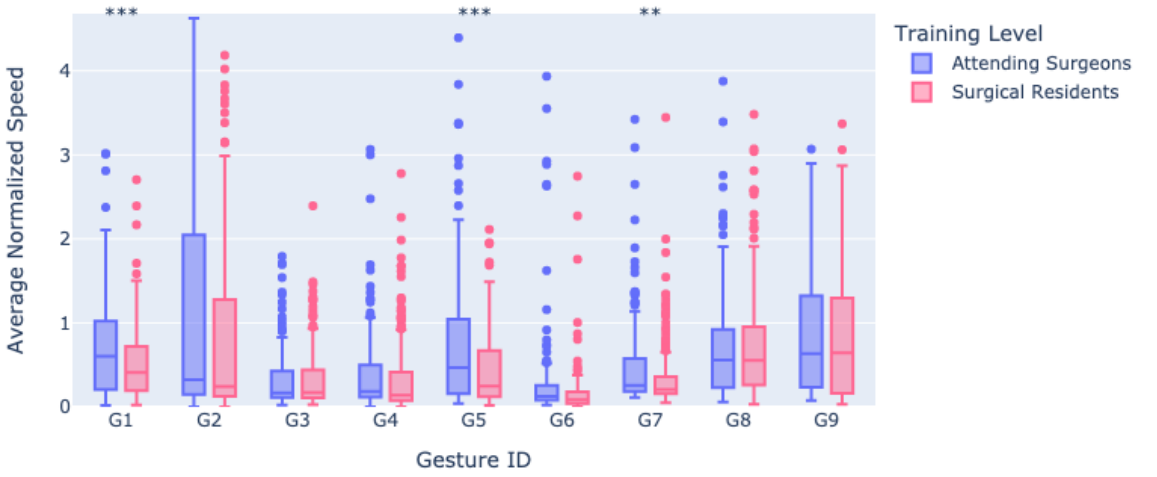


(b) Right grasper's trajectory

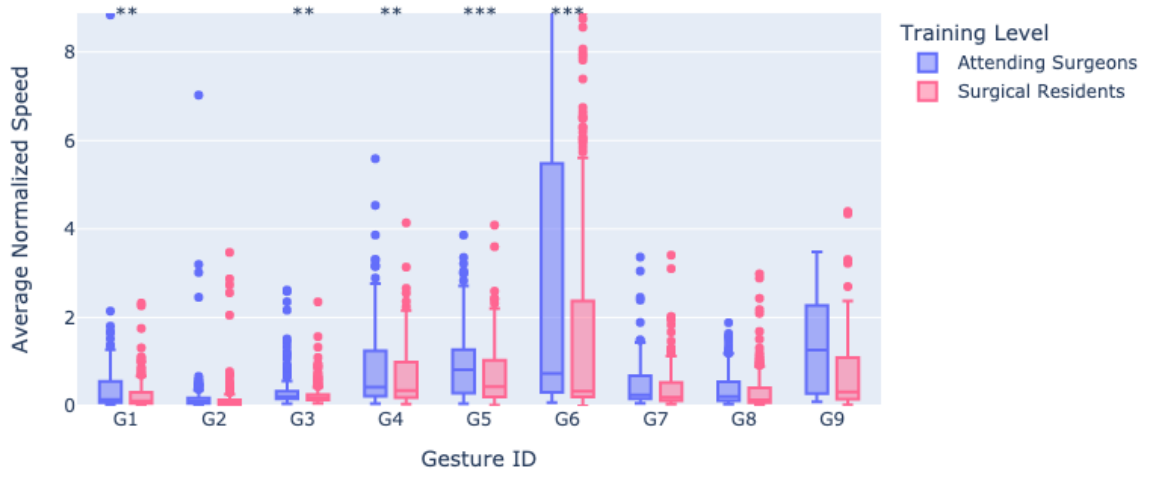
Figure 8.8: Box plots illustrating the normalized path length of the left and right graspers trajectories across the surgical gestures (G1 to G9) during peg transfer procedures, performed by both attending surgeons and surgical residents.

Gesture Speed and Acceleration: Figures 8.9 and 8.10 present box plots showing the average normalized speed and average normalized acceleration across all surgical gestures, respectively. Attending surgeons exibit significantly higher gesture speed for gestures G1, G5, and G7 for the left grasper and G1, G3, G4, G5, and G6 for the right grasper. Conversely, attending surgeons also show significantly higher gesture accelerations for gestures G1, G5, and G7 with the left grasper, and for gestures G1, G3, G4, G5, and G6 with the right grasper.

We notice from these results that there are significant differences in the same surgical gestures across both performance metrics. Specifically, gestures G1, G5, and G7 for the left grasper and G1, G3, G4, G5, and G6 for the right grasper show significant differences in both speed and acceleration metrics. This indicates that gestures performed with greater speed are also those that are performed with the highest acceleration, highlighting a consistent pattern of high-speed and high-acceleration performance by attending surgeons for these



(a) Left grasper's trajectory

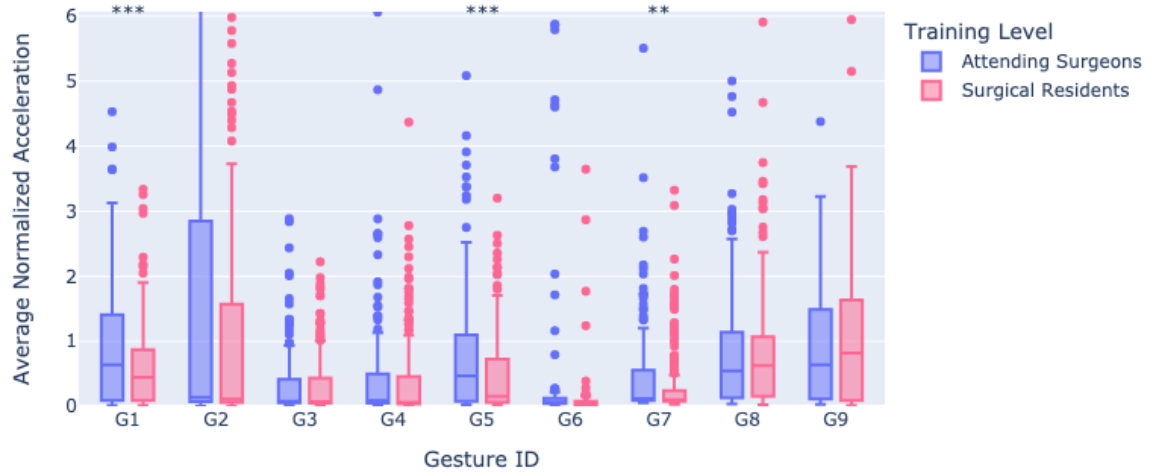


(b) Right grasper's trajectory

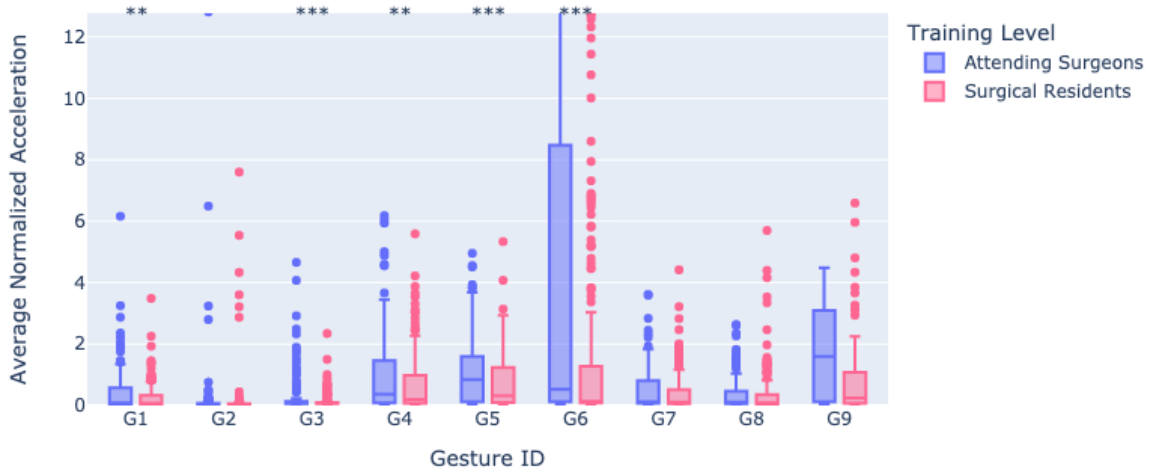
Figure 8.9: Box plots illustrating the averaged normalized speed of the left and right graspers across the different surgical gestures (G1 to G9) during peg transfer procedures, performed by both attending surgeons and surgical residents.

gestures.

Gesture Smoothness: Figure 8.11 highlights the performance metrics for gesture smoothness across different gestures and graspers. Gesture smoothness is quantified by the standard deviation of the jerks, a metric that reflects the variability in acceleration changes during the execution of surgical gestures. Higher values of this metric indicate less smoothness, or greater variability, in the gestures. We can observe that attending surgeons show significantly higher values in gestures G1, G5, and G7 for the left grasper and G3 and G5 for the right grasper.



(a) Left grasper's trajectory



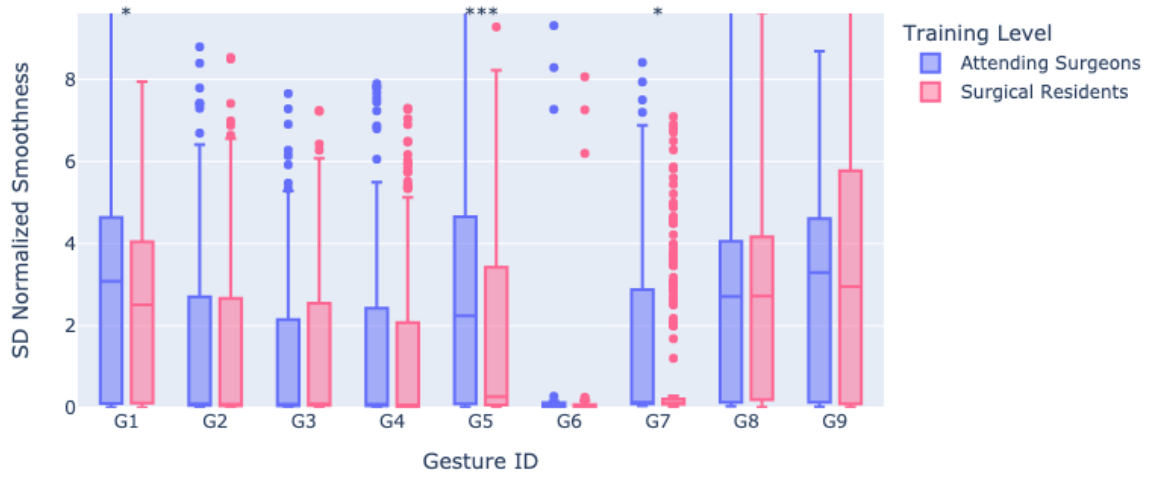
(b) Right grasper's trajectory

Figure 8.10: Box plots illustrating the averaged normalized acceleration of the left and right graspers across the different surgical gestures (G1 to G9) during peg transfer procedures, performed by both attending surgeons and surgical residents.

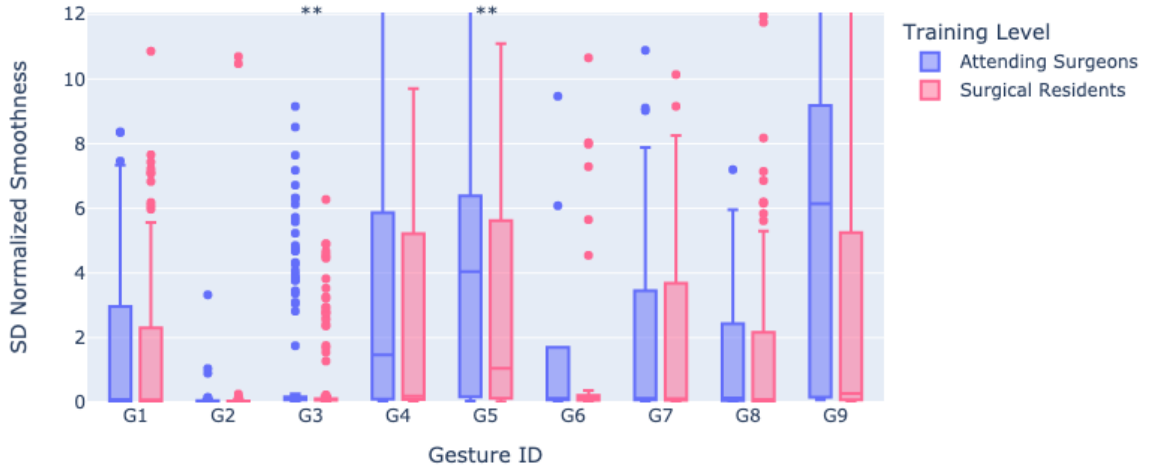
Gesture Curvature: We presented the average normalized curvature in Figure 8.12. For the left grasper, attending surgeons exhibit significantly higher curvature in gestures G1, G3, G4, G7, and G8, suggesting that these gestures involve more pronounced changes in direction, possibly indicating less fluidity or control compared to surgical residents. In contrast, the performance for the right grasper is similar across groups, with no significant differences.

Summary and Implication for Surgical Training

The analysis reveals substantial differences in performance between attending surgeons and surgical residents across various surgical gestures. Attending surgeons complete most



(a) Left grasper's trajectory



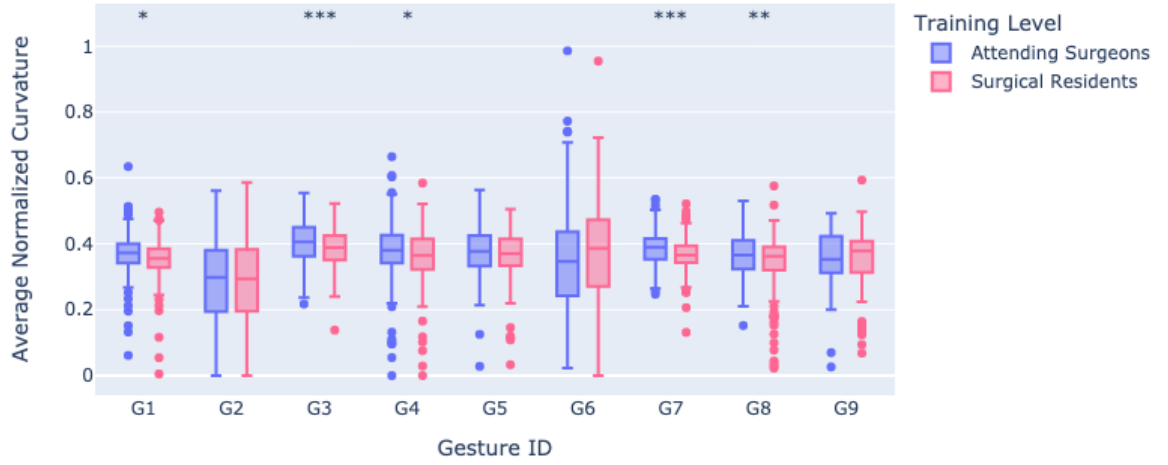
(b) Right grasper's trajectory

Figure 8.11: Box plots illustrating the standard deviation of the gesture smoothness performance metric for the left and right graspers across the different surgical gestures (G1 to G9) during peg transfer procedures, conducted by both attending surgeons and surgical residents.

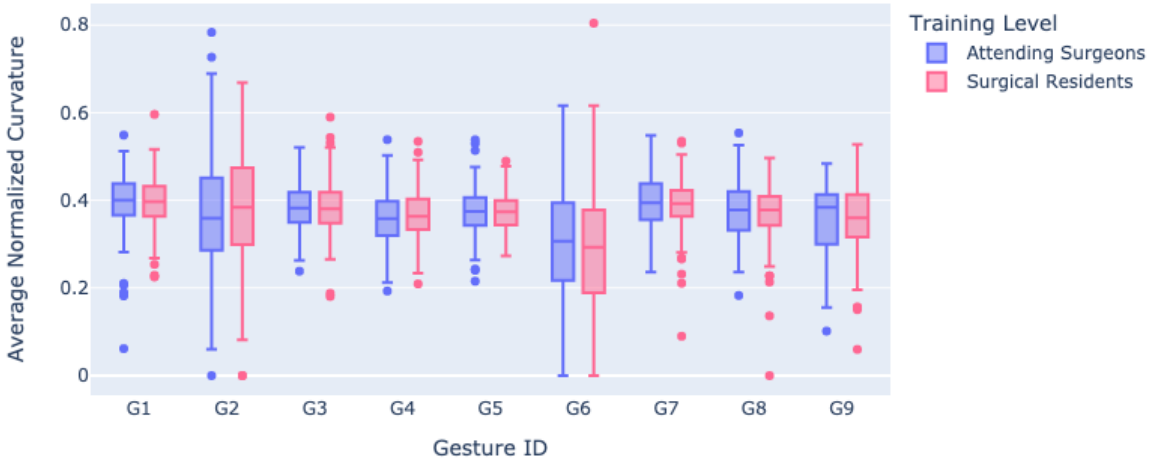
gestures significantly faster, particularly for gestures G1, G2, G3, G4, G5, G7, and G8. In contrast, surgical residents exhibit higher gesture frequency for G3 and G8, indicating that they may rely on increased repetition to accomplish the peg transfer tasks.

Additionally, attending surgeons show higher speeds and accelerations for specific gestures compared to surgical residents. Path length differences are nuanced: attending surgeons have shorter trajectories for gesture G8 with the left grasper, while surgical residents show shorter path lengths for gestures G7 and G6 with the left and right graspers, respectively.

The smoothness analysis, as measured by the standard deviation of the jerk, indicates that attending surgeons exhibit higher variability in gestures G1, G5, and G7 with the left



(a) Left grasper's trajectory



(b) Right grasper's trajectory

Figure 8.12: Box plots illustrating the average normalized curvature for the left and right graspers across the different surgical gestures (G1 to G9) during peg transfer tasks, performed by both attending surgeons and surgical residents.

grasper, and G3 and G5 with the right grasper. Furthermore, the curvature metric reveals that attending surgeons exibit significantly higher curvature in gestures G1, G3, G4, G7, and G8 with the left grasper, suggesting less fluidity.

8.5.2 FPV Suturing

Figures 8.13 and 8.14 depicted the time spent (in seconds) and the frequency of each surgical gestures, respectively, by attending surgeons and medical students during suturing tasks. We can see that attending surgeons performed most surgical gestures in less time. Specifically, significant differences were observed for gestures G0, G1, and G3. Even if there is not a significant difference for G2, we can see in the box plot that medical students have higher Q1, Q3, and maximum values compared to attending surgeons, indicating that medical stu-

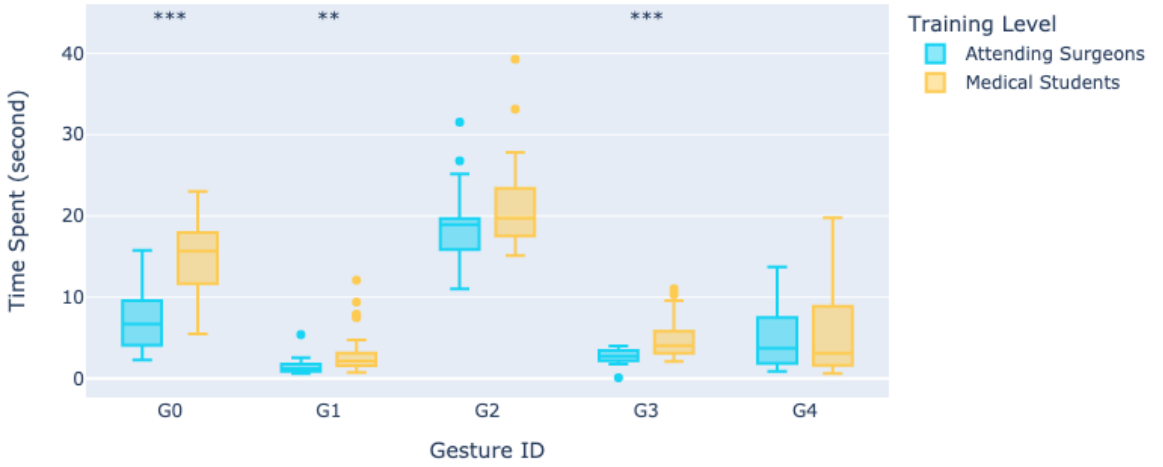


Figure 8.13: Box plots comparing the time spent (in seconds) to complete the different surgical gestures (G0 to G4) during suturing tasks by attending surgeons and medical students. Significance levels are indicated as follows: * p-value < 0.05, ** p-value < 0.01, and *** p-value < 0.001.

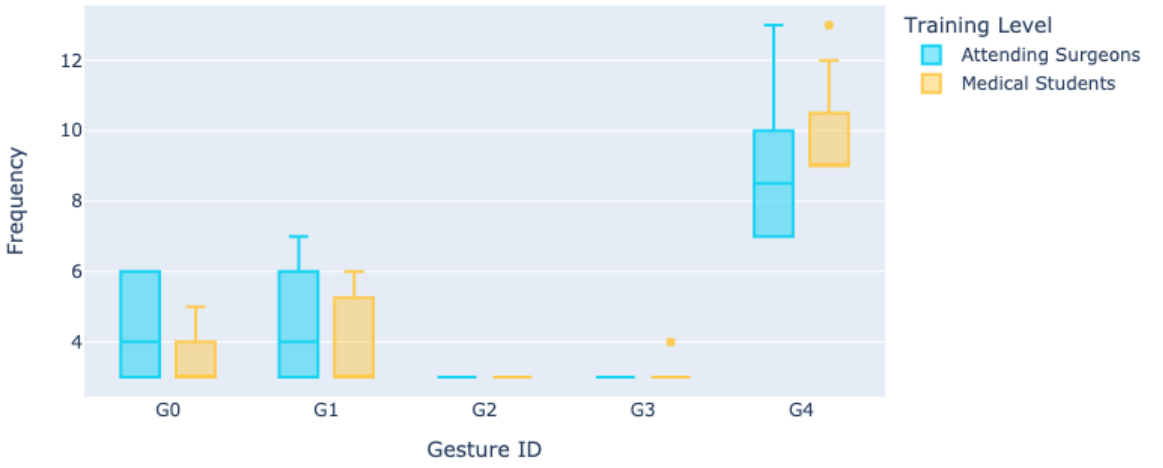


Figure 8.14: Box plots illustrating the frequencies associated with the different surgical gestures (G0 to G4) during suturing procedures, performed by attending surgeons and medical students.

dents tend to perform this gesture with longer duration. The increased time spent on G0 suggests that medical students may experience more delays or inefficiencies during pauses between gestures, potentially impacting the overall fluidity and efficiency of their suturing tasks. Concerning the frequency metric, no significant difference was found between the two groups of practitioners.

8.6 Discussion

8.6.1 Implications of Findings

In this chapter, we introduced two new datasets for the task of surgical gesture recognition: the Peg Transfer dataset and the FPV Suturing dataset. The successful application of these datasets in model training demonstrates their utility and effectiveness as benchmarks.

The Peg Transfer dataset includes both video recordings and surgical tool trajectory data, providing a comprehensive understanding of surgical gestures. The MGRFormer, previously introduced, achieved state-of-the-art performance in the multimodal setting, emphasizing the importance of integrating multiple data sources to enhance recognition accuracy. Specifically, we combined left and right grasper trajectories with video data. In the unimodal benchmark, the ASFormer significantly outperformed other state-of-the-art methods, further establishing the superiority of Transformer-based architectures with refinement modules for surgical gesture recognition.

For the FPV Suturing dataset, which features first-person video recordings of suturing procedures, the ASFormer also achieved state-of-the-art results, demonstrating a substantial margin over existing methods. This dataset's unique perspective offers valuable insights into the subtleties of suturing, opening new avenues for surgical gesture recognition in challenging environments and potential applications in actual surgical operations.

Overall, the strong performance achieved with both datasets, along with their inherent variability—such as differences in positioning and lighting for the Peg Transfer dataset and natural variability in first-person views for the FPV Suturing dataset—indicates that the models trained are well-suited for real-world applications.

8.6.2 Limitations

While the Peg Transfer and FPV Suturing datasets offer significant advancements in surgical gesture recognition, there are notable limitations to address.

One major limitation shared by both datasets is the time-consuming nature of the annotation process. Annotating surgical gestures requires extensive manual effort, which can be both labor-intensive and prone to inconsistencies. This challenge is prevalent across most surgical gesture recognition datasets and can impact the overall efficiency of dataset preparation and model training.

For the FPV Suturing dataset specifically, a notable limitation is the relatively small number of participants, with only five participants contributing to the dataset. Although the ASFormer demonstrated strong performance on this dataset, there remains a performance gap when compared to the Peg Transfer dataset and the VTS dataset. This limitation could be attributed to the small sample size and to the first-person view, which might not fully capture the variability and complexity of suturing gestures.

Additionally, the FPV Suturing dataset involves first-person video recordings that re-

quire participants to wear specialized glasses for data collection. This requirement could pose challenges for real-world applications, where the use of such equipment might not be feasible or practical.

8.6.3 Future Directions

Future research should focus on several key areas to address the limitations identified and further advance the field of surgical gesture recognition.

Firstly, as suggested in the preceding chapter, exploring unsupervised and self-supervised learning approaches could significantly mitigate the challenges associated with the time-consuming and labor-intensive nature of manual annotation. These methods can reduce the reliance on annotated data and improve overall efficiency. Additionally, employing data augmentation techniques could address limitations posed by small dataset sizes, thereby enhancing model performances.

For the FPV Suturing dataset specifically, future work should aim to extend the dataset by including a larger and more diverse set of participants, including both attending surgeons and medical students. This expansion will help to capture a broader range of suturing techniques and variations, ultimately leading to more comprehensive and representative models. Furthermore, augmenting the dataset with additional sources of data, such as tool tracking and hand skeleton tracking, could provide richer contextual information and improve the accuracy of gesture recognition. As we already demonstrated for the Peg Transfer and VTS dataset, the combination of surgical tool tracking or hand movement data with video data has demonstrated superior results compared to unimodal sources of data.

8.7 Conclusion

This chapter introduced two novel datasets designed to address key limitations in existing resources for surgical gesture recognition. The first dataset consists of video recordings of peg transfer tasks, performed by both attending surgeons and surgical residents. For this dataset, the MGRFormer framework outperformed other multimodal approaches by integrating left and right grasper trajectories with video data. Additionally, the ASFormer model achieved superior results in unimodal settings across different data modalities.

The second dataset features first-person perspective (FPV) recordings of suturing tasks performed by both attending surgeons and medical students. Despite challenges such as a limited number of participants and variability in the first-person view, ASFormer delivered strong performance, achieving state-of-the-art results using ResNet-18 and I3D features. This underscores ASFormer's capability to handle complex data modalities, even when faced with visual variability and small sample sizes.

Together, these findings highlight the potential of these datasets and Transformer-based models to advance surgical gesture recognition in diverse and challenging environments. Furthermore, we conducted a comparative statistical analysis of the defined surgical ges-

tures across both datasets, examining performance differences between attending surgeons and medical students. This analysis revealed statistically significant differences in certain performance metrics for specific surgical gestures between the two categories of practitioners. The integration of a surgical gesture recognition system with performance metric computation will enable granular feedback at the gesture level to medical students, enabling targeted recommandations based on their performance.

Chapter 9

Conclusion and Perspectives

9.1 Summary of Contributions

In this thesis, we proposed new deep learning frameworks for stress detection, emotion recognition, surgical skill assessment, and surgical gesture recognition, with the aim of enhancing medical simulation. Firstly, for stress detection, we introduced a multimodal framework that leveraged various fusion techniques to integrate physiological signals from two distinct sensors. Our method outperformed the state-of-the-art approaches for the stress and affect detection tasks. Secondly, for emotion recognition, we introduced the Multimodal Graph-based Transformer framework, which leveraged Graph Convolution Network (GCN) to model the interactions among different levels of modality-specific feature representations extracted using Transformer encoders. Our framework achieved superior performance compared to existing methods, particularly in combining facial landmarks (2D, 3D, and Thermal) with physiological data, facial action units with physiological data, and combining facial landmarks with action units and physiological data. Furthermore, for surgical skill assessment, we proposed using hand skeleton sequences to differentiate between the hand movements of expert and novice practitioners during surgical simulation tasks. To the best of our knowledge, this was the first study to leverage hand skeleton sequences for this purpose. Additionally, we introduced a novel method tailored to this task that combines a GCN and a Transformer encoder. Our approach surpassed state-of-the-art methods across two collected surgical simulation datasets. Lastly, for surgical gesture recognition, we presented the MGRFormer framework, which incorporated an iterative multimodal refinement module designed to enhance the fusion of two different modalities during the refinement stage. Our framework significantly outperformed current methods on a public dataset and is, to the best of our knowledge, the first to explore multimodal fusion at the refinement stage. Lastly, we introduced two new datasets for surgical gesture recognition to address existing dataset limitations. On the collected peg transfer dataset, our MGRFormer outperforms other state-of-the-art methods, further validating its efficacy in leveraging multimodal data for surgical gesture recognition.

9.2 Future Works

While this thesis presents promising advancements in stress detection, emotion recognition, surgical skill assessment, and surgical gesture recognition, several avenues remain unexplored that could further enhance the applicability and generalizability of the proposed frameworks.

9.2.1 Affective Computing

For both stress detection and emotion recognition, our experiments were conducted on existing datasets that may not fully capture the diversity of stress responses and emotional expressions across different populations and environments.

To improve the robustness and generalizability of stress detection and emotion recogni-

tion models, future work should focus on expanding the datasets used for both tasks. In this thesis, the experiments were performed on public datasets, such as WESAD for stress detection and BP4D+ for emotion recognition. However, these datasets may not fully capture the diversity of stress responses and emotional expressions across different populations, demographics, environments, and data collection protocols. Future research should focus on collecting more diverse, larger datasets, capturing spontaneous and naturalistic expressions of stress and emotion in real-world settings. Naturalistic data can provide better insights into how people react in uncontrolled environments.

Moreover, to ensure the generalizability of the developed models, cross-dataset validation is crucial. This process involves training models on one or multiple datasets and then evaluating their performance on completely independent datasets with varying conditions and participant demographics. Cross-dataset validation can reveal how well the model performs beyond the specific context in which it was trained, providing valuable insights into its robustness and applicability across different real-world scenarios.

Furthermore, the subjectivity inherent in labeling stress and emotion presents a significant challenge. Labeling strategies for stress and emotion vary across datasets, from self-reported questionnaires to physiological markers and expert assessments. Future research should focus on establishing more standardized definitions and protocols for both stress and emotion labeling, which could improve consistency across datasets, facilitating more effective cross-dataset validation.

Given the complexity of labeling stress and emotion data and the inherent subjectivity involved, self-supervised learning (SSL) offers a promising approach to enhance stress detection and emotion recognition. By leveraging large amounts of unlabeled physiological and behavioral data, SSL enables models to learn from these datasets through pretext tasks that do not require human annotations. For example, models can learn by reconstructing parts of the input or predicting relationships between different modalities.

One promising direction is the pretraining of our proposed multimodal framework using SSL. For instance, in cross-modal alignment of multimodal data (such as physiological data, facial landmarks, and action units, as seen in the BP4D+ dataset), contrastive learning [230] can be applied to align feature representations from different modalities. The central idea is to bring positive pairs (multimodal data from the same instance or time frame) closer in the feature space, while separating negative pairs (multimodal data from different instances or time frames). More precisely, each modality is processed through a separate network to generate feature embeddings. Positive pairs can be created by pairing physiological data, facial landmarks, and action units from the same time frame or instance (e.g., from the same person experiencing stress or an emotional state). Negative pairs are formed by pairing physiological data from one time frame with facial landmarks or action units from another time frame. A contrastive loss function, such as InfoNCE, is then used to ensure that these embeddings are correctly aligned.

In [231], the authors proposed a multimodal emotion recognition framework that incorporates a modality-pairwise contrastive loss. In their approach, feature representations are first extracted from various modalities, such as text, video, facial landmarks, and acoustic data, using appropriate backbone networks. Subsequently, they computed pairwise con-

trastive loss to make the embeddings of two modalities from the same sequence (positives) closer together, while separating the embeddings of two modalities from different sequences (negatives). The final loss is computed as the sum of the losses obtained from all pairs of modalities.

Lastly, an important direction for future research is the real-world deployment and testing of stress detection and emotion recognition frameworks in medical simulation environments. Integrating these models into training sessions will allow us to capture naturalistic stress and emotional responses from medical students during high-pressure scenarios. The primary goal will be to assess how these technologies enhance medical simulation training, with a focus on evaluating their impact on both student performance and overall learning outcomes.

9.2.2 Surgical Skill Assessment

While our proposed approach for surgical skill assessment using hand skeleton data has demonstrated significant potential, several directions for future work can be pursued to further improve its performance and generalizability.

First, our experiments were conducted on two in-house collected surgical simulation datasets. Future efforts should focus on collecting larger and more diverse datasets, incorporating a greater number of attending surgeons and including multiple trials for each surgical simulation procedure. Additionally, expanding beyond binary classification (expert vs. novice) to a more fine-grained categorization of skill levels could offer deeper insights into the nuances of surgical proficiency.

Transfer learning is another promising avenue for enhancing the performance of surgical skill assessment trained methods. For instance, we could leverage pre-trained models developed for specific surgical tasks as a foundation for training on other simulation tasks, thereby accelerating training and improving generalizability across different procedures. In [232], the authors proposed a Transformer-based framework for action recognition, pre-trained on the NTU RGB+D dataset [233] and fine-tuned on their proposed Tai Chi action recognition dataset. Their experiments demonstrated that this approach achieved high performance even with a small-scale training dataset, suggesting that similar techniques could be effectively applied in the context of surgical skill assessment.

Data augmentation techniques represent another potential area for exploration. Hand skeleton sequences in surgical tasks can exhibit significant variability due to factors such as camera angles, lighting conditions, and individual differences in hand anatomy and motion. To improve the robustness of our model, future work could investigate the use of advanced augmentation techniques, such as geometric transformations, temporal cropping, and synthetic data generation using generative models. This would enable the model to generalize better to across different environments.

Lastly, the real-world deployment of our surgical skill assessment framework in educational settings represents a crucial step for future research. By integrating the model into surgical training simulations, it will be possible to assess how well the system performs

in real-time under practical conditions. Continuous assessment of surgical trainees during simulations could provide invaluable feedback to both instructors and learners, enabling more personalized training programs. Furthermore, long-term studies could evaluate the impact of such technologies on the learning curve and the overall competence of surgical practitioners.

9.2.3 Surgical Gesture Recognition

Firsty, the FPV Suturing dataset introduced in Chapter 8 consists of first-person video recordings from just five participants: two attending surgeons and three medical students performing suturing tasks in a simulated environment. Future research should aim to collect a larger and more diverse dataset. The expanded dataset should include a greater number of participants compared to those in the VTS [222] and Peg Transfer datasets to more accurately capture the variability present in first-person view recordings.

Moreover, as demonstrated by our proposed MGRFormer with the VTS and Peg Transfer datasets, surgical gesture recognition significantly benefits from multimodal learning. Therefore, incorporating an additional modality in the FPV Suturing dataset would enhance the analysis and understanding of surgical gestures. One promising modality could be hand motion tracking or hand skeleton tracking. For instance, Bkheet et al. [234] demonstrated that integrating 2D hand poses with I3D features significantly improved state-of-the-art accuracy in surgical gesture recognition on the VTS dataset. This finding underscores the potential of combining visual features with hand motion data to boost the precision and robustness of gesture recognition systems. Thus, integrating hand motion or hand skeleton tracking into the FPV Suturing dataset could similarly advance the accuracy of surgical gesture analysis, offering a more comprehensive understanding of the nuances involved in surgical techniques.

To address the challenges associated with the labor-intensive and time-consuming process of labeling surgical gestures, future research should explore SSL techniques. These approaches could alleviate some of the bottlenecks in dataset creation by enabling models to learn useful representations from unannotated data. For instance, contrastive learning could be employed to align features from different modalities, thereby improving the model's ability to distinguish between various surgical gestures without extensive manual labeling.

Furthermore, exploring data augmentation techniques could enhance the performance of surgical gesture recognition systems. Methods such as mirroring video recordings, segmenting data sequences into multiple parts, and swapping these segments, followed by training on the modified sequences, can improve model generalizability. These techniques can make the model more resilient to data variations, thereby boosting overall accuracy and robustness in gesture recognition systems.

Another promising avenue is exploring the application of diffusion models for surgical gesture recognition. Recent research has investigated the use of diffusion models in temporal action segmentation. DifffAct, introduced in [235], iteratively refines action segmentation from pure noise, conditioned on video features extracted using an ASFormer.

DifffAct also employs a condition masking strategy that utilizes positional, boundary, and relational priors of human actions to reduce segmentation errors, outperforming the ASFormer in temporal action segmentation tasks. Since the ASFormer was the foundation for developing our MGRFormer, investigating diffusion models for surgical gesture recognition, particularly extensions of DifffAct for integrating multimodal data, could improve the state-of-the-art.

Bibliography

[1] Yujin Wu. "Multimodal emotion recognition from physiological signals and facial expressions". PhD thesis. Université de Lille, 2023.

[2] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[3] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[4] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[5] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[6] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[7] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[8] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[9] Linhao Dong, Shuang Xu, and Bo Xu. "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5884–5888.

[10] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[11] Emilio Parisotto et al. "Stabilizing transformers for reinforcement learning". In: *International conference on machine learning*. PMLR. 2020, pp. 7487–7498.

[12] Haoyi Zhou et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 11106–11115.

[13] Jiasen Lu et al. “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks”. In: *Advances in neural information processing systems* 32 (2019).

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).

[15] Jonathan Godwin et al. “Simple gnn regularisation for 3d molecular property prediction & beyond”. In: *arXiv preprint arXiv:2106.07971* (2021).

[16] Hanxuan Cai et al. “FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction”. In: *Briefings in bioinformatics* 23.6 (2022), bbac408.

[17] Oliver Wieder et al. “A compact review of molecular property prediction with graph neural networks”. In: *Drug Discovery Today: Technologies* 37 (2020), pp. 1–12.

[18] Zhiwei Guo and Heng Wang. “A deep graph neural network-based mechanism for social recommendations”. In: *IEEE Transactions on Industrial Informatics* 17.4 (2020), pp. 2776–2783.

[19] Shengjie Min et al. “Stgsn—a spatial–temporal graph neural network framework for time-evolving social networks”. In: *Knowledge-Based Systems* 214 (2021), p. 106746.

[20] Sanjay Kumar et al. “Influence maximization in social networks using graph embedding and graph neural network”. In: *Information Sciences* 607 (2022), pp. 1617–1636.

[21] Yue Wang et al. “Dynamic graph cnn for learning on point clouds”. In: *ACM Transactions on Graphics (tog)* 38.5 (2019), pp. 1–12.

[22] Kai Han et al. “Vision gnn: An image is worth graph of nodes”. In: *Advances in neural information processing systems* 35 (2022), pp. 8291–8303.

[23] Osman Boyaci et al. “Graph neural networks based detection of stealth false data injection attacks in smart grids”. In: *IEEE Systems Journal* 16.2 (2021), pp. 2946–2957.

[24] Seyed Hamed Haghshenas, Md Abul Hasnat, and Mia Naeini. “A temporal graph neural network for cyber attack detection and localization in smart grids”. In: *2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE. 2023, pp. 1–5.

[25] Mohammad Soleymani, Maja Pantic, and Thierry Pun. “Multimodal emotion recognition in response to videos”. In: *IEEE transactions on affective computing* 3.2 (2011), pp. 211–223.

[26] Abhinav Joshi et al. “COGMEN: COntextualized GNN based multimodal emotion recognitioN”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, pp. 4148–4164.

[27] Stanislaw Antol et al. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[28] Jiasen Lu et al. "Hierarchical question-image co-attention for visual question answering". In: *Advances in neural information processing systems* 29 (2016).

[29] Yidan Qin et al. "Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 371–377.

[30] Yonghao Long et al. "Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 13346–13353.

[31] Beatrice Van Amsterdam et al. "Gesture recognition in robotic surgery with multimodal attention". In: *IEEE Transactions on Medical Imaging* 41.7 (2022), pp. 1677–1687.

[32] Harold Hotelling. "Relations between two sets of variates". In: *Breakthroughs in statistics: methodology and distribution*. Springer, 1992, pp. 162–190.

[33] Galen Andrew et al. "Deep canonical correlation analysis". In: *International conference on machine learning*. PMLR. 2013, pp. 1247–1255.

[34] Yao-Hung Hubert Tsai et al. "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for computational linguistics. Meeting*. Vol. 2019. NIH Public Access. 2019, p. 6558.

[35] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[36] Avrim Blum and Tom Mitchell. "Combining labeled and unlabeled data with co-training". In: *Proceedings of the eleventh annual conference on Computational learning theory*. 1998, pp. 92–100.

[37] Fuzhen Zhuang et al. "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[38] Rich Caruana. "Multitask learning". In: *Machine learning* 28 (1997), pp. 41–75.

[39] Amir Zadeh et al. "Tensor fusion network for multimodal sentiment analysis". In: *arXiv preprint arXiv:1707.07250* (2017).

[40] Xueming Yan et al. "Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling". In: *Applied Artificial Intelligence* 36.1 (2022), p. 2000688.

[41] Chiori Hori et al. "Attention-based multimodal fusion for video description". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4193–4202.

[42] Yuanchao Li, Tianyu Zhao, and Xun Shen. "Attention-based multimodal fusion for estimating human emotion in real-world HRI". In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020, pp. 340–342.

[43] Jingwen Hu et al. "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation". In: *arXiv preprint arXiv:2107.06779* (2021).

[44] David Matsumoto. "Culture and emotion". In: *The handbook of culture and psychology* (2001), pp. 171–194.

[45] Thomas J Bouchard Jr. "Genes, environment, and personality". In: *Science* 264.5166 (1994), pp. 1700–1701.

[46] Paul Ekman and Wallace V Friesen. "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2 (1971), p. 124.

[47] Paul Ekman. "An argument for basic emotions". In: *Cognition & emotion* 6.3-4 (1992), pp. 169–200.

[48] Paul Ekman and Wallace V Friesen. "Facial action coding system". In: *Environmental Psychology & Nonverbal Behavior* (1978).

[49] Daniel T Cordaro et al. "The voice conveys emotion in ten globalized cultures and one remote village in Bhutan." In: *Emotion* 16.1 (2016), p. 117.

[50] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[51] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. "Deep region and multi-label learning for facial action unit detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3391–3399.

[52] Geethu Miriam Jacob and Bjorn Stenger. "Facial action unit detection with transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 7680–7689.

[53] Michel Valstar and Maja Pantic. "Fully automatic facial action unit detection and temporal analysis". In: *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE. 2006, pp. 149–149.

[54] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. "Detector of facial landmarks learned by the structured output SVM". In: *International conference on computer vision theory and applications*. Vol. 2. SCITEPRESS. 2012, pp. 547–556.

[55] Yue Wu et al. "Facial landmark detection with tweaked convolutional neural networks". In: *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017), pp. 3067–3074.

[56] Zhanpeng Zhang et al. "Facial landmark detection by deep multi-task learning". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer. 2014, pp. 94–108.

[57] Zhe Cao et al. “Realtime multi-person 2d pose estimation using part affinity fields”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299.

[58] Laurie Kelly McCorry. “Physiology of the autonomic nervous system”. In: *American journal of pharmaceutical education* 71.4 (2007).

[59] Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012.

[60] Xiaorong Pu et al. “Facial expression recognition from image sequences using twofold random forest classifier”. In: *Neurocomputing* 168 (2015), pp. 1173–1180.

[61] Patrick Lucey et al. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression”. In: *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE. 2010, pp. 94–101.

[62] Pranav Kumar, SL Happy, and Aurobinda Routray. “A real-time robust facial expression recognition system using HOG features”. In: *2016 International Conference on Computing, Analytics and Security Trends (CAST)*. IEEE. 2016, pp. 289–293.

[63] Muzammil Abdulrahman and Alaa Eleyan. “Facial expression recognition using support vector machines”. In: *2015 23nd signal processing and communications applications conference (SIU)*. IEEE. 2015, pp. 276–279.

[64] Sébastien Ouellet. “Real-time emotion recognition for gaming using deep convolutional network features”. In: *arXiv preprint arXiv:1408.3750* (2014).

[65] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[66] Bin Li and Dimas Lima. “Facial expression recognition via ResNet-50”. In: *International Journal of Cognitive Computing in Engineering* 2 (2021), pp. 57–64.

[67] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[68] Yin Fan et al. “Video-based emotion recognition using CNN-RNN and C3D hybrid networks”. In: *Proceedings of the 18th ACM international conference on multimodal interaction*. 2016, pp. 445–450.

[69] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[70] Ce Zheng, Matias Mendieta, and Chen Chen. “Poster: A pyramid cross-fusion transformer network for facial expression recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3146–3155.

[71] Cornelia Setz et al. “Discriminating stress from cognitive load using a wearable EDA device”. In: *IEEE Transactions on information technology in biomedicine* 14.2 (2009), pp. 410–417.

[72] Martin Ragot et al. “Emotion recognition using physiological signals: laboratory vs. wearable sensors”. In: *Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*. Springer. 2018, pp. 15–22.

[73] Wanhui Wen et al. “Emotion recognition based on multi-variant correlation of physiological signals”. In: *IEEE Transactions on Affective Computing* 5.2 (2014), pp. 126–140.

[74] Yu-Liang Hsu et al. “Automatic ECG-based emotion recognition in music listening”. In: *IEEE Transactions on Affective Computing* 11.1 (2017), pp. 85–99.

[75] Terumi Umematsu et al. “Improving students’ daily life stress forecasting using LSTM neural networks”. In: *2019 IEEE EMBS international conference on biomedical & health informatics (BHI)*. IEEE. 2019, pp. 1–4.

[76] Muhammad Awais et al. “LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19”. In: *IEEE Internet of Things Journal* 8.23 (2020), pp. 16863–16871.

[77] Muhammad Najam Dar et al. “CNN and LSTM-based emotion charting using physiological signals”. In: *Sensors* 20.16 (2020), p. 4551.

[78] Tomasz Wierciński et al. “Emotion recognition from physiological channels using graph neural network”. In: *Sensors* 22.8 (2022), p. 2980.

[79] Juan Abdon Miranda-Correa et al. “Amigos: A dataset for affect, personality and mood research on individuals and groups”. In: *IEEE transactions on affective computing* 12.2 (2018), pp. 479–493.

[80] Juan Vazquez-Rodriguez et al. “Transformer-based self-supervised learning for emotion recognition”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022, pp. 2605–2612.

[81] Dahua Li et al. “Facial expression recognition based on electroencephalogram and facial landmark localization”. In: *Technology and Health Care* 27.4 (2019), pp. 373–387.

[82] Nastaran Saffaryazdi et al. “Using facial micro-expressions in combination with EEG and physiological signals for emotion recognition”. In: *Frontiers in Psychology* 13 (2022), p. 864047.

[83] Sander Koelstra et al. “Deap: A database for emotion analysis; using physiological signals”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[84] Longfei Yang et al. “The effects of psychological stress on depression”. In: *Current neuropharmacology* 13.4 (2015), pp. 494–504.

[85] W. B. Cannon. *The Wisdom of the Body*. PDF. Norton, Massachusetts: The Norton Library, 1932.

[86] Alison N Saul et al. "Chronic stress and susceptibility to skin cancer". In: *Journal of the National Cancer Institute* 97.23 (2005), pp. 1760–1767.

[87] Ronald Glaser and Janice K Kiecolt-Glaser. "Stress-induced immune dysfunction: implications for health". In: *Nature Reviews Immunology* 5.3 (2005), pp. 243–251.

[88] Sheldon Cohen, David AJ Tyrrell, and Andrew P Smith. "Psychological stress and susceptibility to the common cold". In: *New England journal of medicine* 325.9 (1991), pp. 606–612.

[89] Giorgos Giannakakis et al. "Review on psychological stress detection using biosignals". In: *IEEE transactions on affective computing* 13.1 (2019), pp. 440–460.

[90] Philip Schmidt et al. "Introducing wesad, a multimodal dataset for wearable stress and affect detection". In: *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018, pp. 400–408.

[91] Sirat Samyoun, Abu Sayeed Mondol, and John A Stankovic. "Stress detection via sensor translation". In: *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE. 2020, pp. 19–26.

[92] Shiyang Li et al. "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting". In: *Advances in neural information processing systems* 32 (2019).

[93] Neo Wu et al. "Deep transformer models for time series forecasting: The influenza prevalence case". In: *arXiv preprint arXiv:2001.08317* (2020).

[94] George Zerveas et al. "A transformer-based framework for multivariate time series representation learning". In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 2114–2124.

[95] Minghao Liu et al. "Gated transformer networks for multivariate time series classification". In: *arXiv preprint arXiv:2103.14438* (2021).

[96] Yuan Yuan and Lei Lin. "Self-supervised pretraining of transformers for satellite image time series classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), pp. 474–487.

[97] Jiehui Xu et al. "Anomaly transformer: Time series anomaly detection with association discrepancy". In: *arXiv preprint arXiv:2110.02642* (2021).

[98] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. "Tranad: Deep transformer networks for anomaly detection in multivariate time series data". In: *arXiv preprint arXiv:2201.07284* (2022).

[99] Zekai Chen et al. "Learning graph structures with transformer for multivariate time-series anomaly detection in IoT". In: *IEEE Internet of Things Journal* 9.12 (2021), pp. 9179–9189.

[100] Jian Huang et al. "Multimodal transformer fusion for continuous emotion recognition". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3507–3511.

[101] Swalpa Kumar Roy et al. “Multimodal fusion transformer for remote sensing image classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–20.

[102] Arsha Nagrani et al. “Attention bottlenecks for multimodal fusion”. In: *Advances in neural information processing systems* 34 (2021), pp. 14200–14213.

[103] Lam Huynh et al. “Stressnas: Affect state and stress detection using neural architecture search”. In: *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 2021, pp. 121–125.

[104] Nafiul Rashid et al. “Feature augmented hybrid cnn for stress recognition using wrist-based photoplethysmography sensor”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 2374–2377.

[105] Yujin WU et al. “Fusion of Physiological and Behavioural Signals on SPD Manifolds with Application to Stress and Pain Detection”. In: *arXiv preprint arXiv:2207.08811* (2022).

[106] Jonathan Aigrain et al. “Multimodal stress detection from multiple assessments”. In: *IEEE Transactions on Affective Computing* 9.4 (2016), pp. 491–506.

[107] Luntian Mou et al. “Driver stress detection via multimodal fusion using attention-based CNN-LSTM”. In: *Expert Systems with Applications* 173 (2021), p. 114693.

[108] Empatica. *Empatica E4 wristband*. Accessed: 2024-05-02. 2024. URL: https://www.empatica.com/en-eu/research/e4/.

[109] RespiBAN. *RespiBAN*. Accessed: 2021-08-08. 2021. URL: https://plux.info/biosignalsplux-wearables/313-respiban-professional-820202407.html.

[110] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. “The ‘Trier Social Stress Test’–a tool for investigating psychobiological stress responses in a laboratory setting”. In: *Neuropsychobiology* 28.1-2 (1993), pp. 76–81.

[111] Manuel Gil-Martin et al. “Human stress detection with wearable sensors using convolutional neural networks”. In: *IEEE Aerospace and Electronic Systems Magazine* 37.1 (2022), pp. 60–70.

[112] Kenneth Lai, Svetlana N Yanushkevich, and Vlad P Shmerko. “Intelligent stress monitoring assistant for first responders”. In: *IEEE Access* 9 (2021), pp. 25314–25329.

[113] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[114] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[115] Yang Wu et al. "Leveraging Multi-modal Interactions among the Intermediate Representations of Deep Transformers for Emotion Recognition". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 2022, pp. 101–109.

[116] Andy T Liu et al. "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders". In: *IEEE ICASSP*. 2020, pp. 6419–6423.

[117] Hamid Reza Vaezi Joze et al. "MMTM: Multimodal transfer module for CNN fusion". In: *IEEE CVPR*. 2020, pp. 13289–13299.

[118] Hassan Akbari et al. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24206–24221.

[119] Zheng Zhang et al. "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[120] Patrick Lucey et al. "Painful data: The UNBC-McMaster shoulder pain expression archive database". In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE. 2011, pp. 57–64.

[121] Rizwan Ahmed Khan et al. "Pain detection through shape and appearance features". In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2013, pp. 1–6.

[122] Ahmed Ashraf, Anqi Yang, and Babak Taati. "Pain expression recognition using occluded faces". In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE. 2019, pp. 1–5.

[123] Corneliu Florea, Laura Florea, and Constantin Vertan. "Learning pain from emotion: transferred hot data representation for pain intensity estimation". In: *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*. Springer. 2015, pp. 778–790.

[124] Ghada Zamzmi et al. "Convolutional neural networks for neonatal pain assessment". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1.3 (2019), pp. 192–200.

[125] Y Zarghami et al. "Pain Detection in Masked Faces during Procedural Sedation". In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2023, pp. 1–6.

[126] Saining Xie et al. "Aggregated residual transformations for deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500.

[127] Safaa El Morabit and Atika Rivenq. "Pain detection from facial expressions based on transformers and distillation". In: *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE. 2022, pp. 1–5.

[128] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International conference on machine learning*. PMLR. 2021, pp. 10347–10357.

[129] Xin Yuan et al. “Occluded Facial Pain Assessment in the ICU using Action Units Guided Network”. In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[130] Camillo Lugaresi et al. “Mediapipe: A framework for building perception pipelines”. In: *arXiv preprint arXiv:1906.08172* (2019).

[131] Benjamin Szczapa et al. “Automatic estimation of self-reported pain by interpretable representations of motion dynamics”. In: *IEEE ICPR*. 2021, pp. 2544–2550.

[132] Antoni Mauricio et al. “A sequential approach for pain recognition based on facial representations”. In: *Computer Vision Systems: 12th International Conference, ICVS 2019, Thessaloniki, Greece, September 23–25, 2019, Proceedings 12*. Springer. 2019, pp. 295–304.

[133] Ghazal Bargshady et al. “Enhanced deep learning algorithm development to detect pain intensity from facial expression images”. In: *Expert systems with applications* 149 (2020), p. 113305.

[134] Ghazal Bargshady et al. “A joint deep neural network model for pain recognition from face”. In: *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. IEEE. 2019, pp. 52–56.

[135] Ghazal Bargshady et al. “Ensemble neural network approach detecting pain intensity from facial expressions”. In: *Artificial Intelligence in Medicine* 109 (2020), p. 101954.

[136] Haochen Xu and Manhua Liu. “A deep attention transformer network for pain estimation with facial expression video”. In: *Biometric Recognition: 15th Chinese Conference, CCBR 2021, Shanghai, China, September 10–12, 2021, Proceedings 15*. Springer. 2021, pp. 112–119.

[137] Busra T Susam et al. “Automated pain assessment using electrodermal activity data and machine learning”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 372–375.

[138] Yaqi Chu et al. “Physiological signals based quantitative evaluation method of the pain”. In: *IFAC Proceedings Volumes* 47.3 (2014), pp. 2981–2986.

[139] Yaqi Chu et al. “Physiological signal-based method for measurement of pain intensity”. In: *Frontiers in neuroscience* 11 (2017), p. 279.

[140] Mingxin Yu et al. “Diverse frequency band-based convolutional neural networks for tonic cold pain assessment using EEG”. In: *Neurocomputing* 378 (2020), pp. 270–282.

[141] Jiahao Wang et al. “An autoencoder-based approach to predict subjective pain perception from high-density evoked EEG potentials”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 1507–1511.

[142] Tor D Wager et al. "An fMRI-based neurologic signature of physical pain". In: *New England Journal of Medicine* 368.15 (2013), pp. 1388–1397.

[143] Ahmad Pourshoghi, Issa Zakeri, and Kambiz Pourrezaei. "Application of functional data analysis in classification and clustering of functional near-infrared spectroscopy signal in response to noxious stimuli". In: *Journal of Biomedical Optics* 21.10 (2016), pp. 101411–101411.

[144] Ruicong Zhi et al. "Multimodal-based stream integrated neural networks for pain assessment". In: *IEICE TRANSACTIONS on Information and Systems* 104.12 (2021), pp. 2184–2194.

[145] Md Sirajus Salekin et al. "Multi-channel neural network for assessing neonatal pain from videos". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE. 2019, pp. 1551–1556.

[146] Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[147] Md Sirajus Salekin et al. "Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment". In: *Computers in biology and medicine* 129 (2021), p. 104150.

[148] Saurabh Hinduja, Shaun Canavan, and Gurmeet Kaur. "Multimodal fusion of physiological signals and facial action units for pain recognition". In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE. 2020, pp. 577–581.

[149] Philipp Werner et al. "Automatic recognition methods supporting pain assessment: A survey". In: *IEEE Transactions on Affective Computing* 13.1 (2019), pp. 530–552.

[150] Stefanos Gkikas and Manolis Tsiknakis. "Automatic assessment of pain based on deep learning methods: A systematic review". In: *Computer methods and programs in biomedicine* 231 (2023), p. 107365.

[151] Matthias Fey and Jan Eric Lenssen. "Fast graph representation learning with PyTorch Geometric". In: *arXiv preprint arXiv:1903.02428* (2019).

[152] Yibo Huang et al. "HybNet: a hybrid network structure for pain intensity estimation". In: *The Visual Computer* 38.3 (2022), pp. 871–882.

[153] Keng Wah Choo and Tiehua Du. "Pain detection from facial landmarks using spatial-temporal deep neural network". In: *ICDIP 2021*. Vol. 11878. SPIE, pp. 593–597.

[154] JA Martin et al. "Objective structured assessment of technical skill (OSATS) for surgical residents". In: *British journal of surgery* 84.2 (1997), pp. 273–278.

[155] Alvin C Goh et al. "Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills". In: *The Journal of urology* 187.1 (2012), pp. 247–252.

[156] Bing Yu, Haoteng Yin, and Zhanxing Zhu. "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting". In: *arXiv preprint arXiv:1709.04875* (2017).

[157] Shengnan Guo et al. "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 2019, pp. 922–929.

[158] Minbo Ma et al. "Histgnn: Hierarchical spatio-temporal graph neural network for weather forecasting". In: *Information Sciences* 648 (2023), p. 119580.

[159] Ke Cheng et al. "Skeleton-based action recognition with shift graph convolutional network". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 183–192.

[160] Rim Slama, Wael Rabah, and Hazem Wannous. "STr-GCN: Dual Spatial Graph Convolutional Network and Transformer Graph Encoder for 3D Hand Gesture Recognition". In: *IEEE FG*. 2023, pp. 1–6.

[161] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. "Spatial temporal transformer network for skeleton-based action recognition". In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*. Springer. 2021, pp. 694–701.

[162] Yuhan Zhang et al. "STST: Spatial-temporal specialized transformer for skeleton-based action recognition". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 3229–3237.

[163] Vittorio Mazzia et al. "Action Transformer: A self-attention model for short-time pose-based human action recognition". In: *Pattern Recognition* 124 (2022), p. 108487.

[164] Yixin Gao et al. "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling". In: *MICCAI workshop: M2cai*. Vol. 3. 3. 2014.

[165] Lingling Tao et al. "Sparse hidden markov models for surgical gesture classification and skill evaluation". In: *Information Processing in Computer-Assisted Interventions: Third International Conference, IPCAI 2012, Pisa, Italy, June 27, 2012. Proceedings 3*. Springer. 2012, pp. 167–177.

[166] Mahtab J Fard et al. "Automated robot-assisted surgical skill evaluation: Predictive analytics approach". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 14.1 (2018), e1850.

[167] Aneeq Zia and Irfan Essa. "Automated surgical skill assessment in RMIS training". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 731–739.

[168] Ziheng Wang and Ann Majewicz Fey. "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 1959–1970.

[169] Hassan Ismail Fawaz et al. "Evaluating surgical skills from kinematic data using convolutional neural networks". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 214–221.

[170] Dayvid Castro et al. "Towards optimizing convolutional neural networks for robotic surgery skill evaluation". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.

[171] Ziheng Wang and Ann Majewicz Fey. "SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 1793–1796.

[172] Xuan Anh Nguyen et al. "Surgical skill levels: Classification and analysis using deep neural network model and motion signals". In: *Computer methods and programs in biomedicine* 177 (2019), pp. 1–8.

[173] Aneeq Zia et al. "Video and accelerometer-based motion analysis for automated surgical skills assessment". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 443–455.

[174] Sanchit Hira et al. "Video-based assessment of intraoperative surgical skill". In: *International journal of computer assisted radiology and surgery* 17.10 (2022), pp. 1801–1811.

[175] Isabel Funke et al. "Video-based surgical skill assessment using 3D convolutional neural networks". In: *International journal of computer assisted radiology and surgery* 14 (2019), pp. 1217–1225.

[176] Limin Wang et al. "Temporal segment networks for action recognition in videos". In: *IEEE transactions on pattern analysis and machine intelligence* 41.11 (2018), pp. 2740–2755.

[177] Daochang Liu et al. "Towards unified surgical skill assessment". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9522–9531.

[178] Fan Zhang et al. "Mediapipe hands: On-device real-time hand tracking". In: *arXiv preprint arXiv:2006.10214* (2020).

[179] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. "Skeleton-based dynamic hand gesture recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 1–9.

[180] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling". In: *arXiv preprint arXiv:1803.01271* (2018).

[181] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[182] Mehran Maghoumi and Joseph J LaViola. "DeepGRU: Deep gesture recognition utility". In: *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14*. Springer. 2019, pp. 16–31.

[183] Michele Tonutti et al. "The role of technology in minimally invasive surgery: state of the art, recent developments and future directions". In: *Postgraduate medical journal* 93.1097 (2017), pp. 159–167.

[184] Ryan W Dobbs et al. "Single-port robotic surgery: the next generation of minimally invasive urology". In: *World journal of urology* 38 (2020), pp. 897–905.

[185] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. "Asformer: Transformer for action segmentation". In: *arXiv preprint arXiv:2110.08568* (2021).

[186] Jiahui Wang et al. "Cross-enhancement transformer for action segmentation". In: *Multimedia Tools and Applications* 83.9 (2024), pp. 25643–25656.

[187] Xiaoyan Tian, Ye Jin, and Xianglong Tang. "Local–global transformer neural network for temporal action segmentation". In: *Multimedia Systems* 29.2 (2023), pp. 615–626.

[188] Adam Goldbraikh et al. "Using open surgery simulation kinematic data for tool and gesture recognition". In: *International Journal of Computer Assisted Radiology and Surgery* 17.6 (2022), pp. 965–979.

[189] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. "Fast saliency based pooling of fisher encoded dense trajectories". In: *ECCV THUMOS Workshop*. Vol. 1. 2. 2014, p. 5.

[190] Marcus Rohrbach et al. "A database for fine grained activity detection of cooking activities". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 1194–1201.

[191] Yu Cheng et al. "Temporal sequence modeling for video event detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 2227–2234.

[192] Hamed Pirsiavash and Deva Ramanan. "Parsing videos of actions with segmental grammars". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 612–619.

[193] Lingling Tao et al. "Surgical gesture segmentation and recognition". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part III 16*. Springer. 2013, pp. 339–346.

[194] Hilde Kuehne, Juergen Gall, and Thomas Serre. "An end-to-end generative framework for video segmentation and recognition". In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, pp. 1–8.

[195] Bharat Singh et al. "A multi-stream bi-directional recurrent neural network for fine-grained action detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1961–1970.

[196] Jeffrey Donahue et al. "Long-term recurrent convolutional networks for visual recognition and description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634.

[197] Colin Lea et al. "Temporal convolutional networks for action segmentation and detection". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 156–165.

[198] Peng Lei and Sinisa Todorovic. "Temporal deformable residual networks for action segmentation in videos". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6742–6751.

[199] Yazan Abu Farha and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3575–3584.

[200] Shi-Jie Li et al. "Ms-tcn++: Multi-stage temporal convolutional network for action segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* (2020).

[201] Yifei Huang, Yusuke Sugano, and Yoichi Sato. "Improving action segmentation via graph-based temporal reasoning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 14024–14034.

[202] Junbin Zhang, Pei-Hsuan Tsai, and Meng-Hsun Tsai. "Semantic2graph: Graph-based multi-modal feature for action segmentation in videos". In: *arXiv preprint arXiv:2209.05653* (2022).

[203] Jiahui Wang et al. "Cross-enhancement transformer for action segmentation". In: *Multimedia Tools and Applications* (2023), pp. 1–14.

[204] Dazhao Du et al. "Do We Really Need Temporal Convolutions in Action Segmentation?" In: *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2023, pp. 1014–1019.

[205] Zexing Du and Qing Wang. "Dilated transformer with feature aggregation module for action segmentation". In: *Neural Processing Letters* (2022), pp. 1–17.

[206] Balakrishnan Varadarajan et al. "Data-derived models for segmentation with application to surgical assessment and training". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009: 12th International Conference, London, UK, September 20-24, 2009, Proceedings, Part I 12*. Springer. 2009, pp. 426–434.

[207] Hassan Ismail Fawaz et al. "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks". In: *International journal of computer assisted radiology and surgery* 14 (2019), pp. 1611–1617.

[208] Colin Lea et al. “Temporal convolutional networks: A unified approach to action segmentation”. In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 47–54.

[209] Robert DiPietro et al. “Recognizing surgical activities with recurrent neural networks”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19*. Springer. 2016, pp. 551–558.

[210] Robert DiPietro et al. “Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks”. In: *International journal of computer assisted radiology and surgery* 14.11 (2019), pp. 2005–2020.

[211] Chang Shi, Yi Zheng, and Ann Majewicz Fey. “Recognition and Prediction of Surgical Gestures and Trajectories Using Transformer Models in Robot-Assisted Surgery”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 8017–8024.

[212] Isabel Funke et al. “Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, pp. 467–475.

[213] Jinglu Zhang et al. “Symmetric dilated convolution for surgical gesture recognition”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 409–418.

[214] Daochang Liu and Tingting Jiang. “Deep reinforcement learning for surgical gesture segmentation and classification”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 247–255.

[215] Colin Lea, Gregory D Hager, and Rene Vidal. “An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks”. In: *2015 IEEE winter conference on applications of computer vision*. IEEE. 2015, pp. 1123–1129.

[216] Adithyavairavan Murali et al. “Tsc-dl: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning”. In: *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2016, pp. 4150–4157.

[217] Yidan Qin et al. “davincinet: Joint prediction of motion and surgical state in robot-assisted surgery”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 2921–2928.

[218] Jian-Fang Hu et al. “Deep bilinear learning for rgb-d action recognition”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 335–351.

[219] Benjia Zhou et al. “Decoupling and recoupling spatiotemporal representation for RGB-D-based motion recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20154–20163.

[220] Benjia Zhou et al. “A Unified Multimodal De-and Re-coupling Framework for RGB-D Motion Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[221] Joao Carreira and Andrew Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[222] Adam Goldbraikh et al. “Video-based fully automatic assessment of open surgery suturing skills”. In: *International Journal of Computer Assisted Radiology and Surgery* 17.3 (2022), pp. 437–448.

[223] Anne-Lise D D’Angelo et al. “Idle time: an underdeveloped performance metric for assessing surgical skill”. In: *The American journal of surgery* 209.4 (2015), pp. 645–651.

[224] E Matt Ritter and Daniel J Scott. “Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery”. In: *Surgical innovation* 14.2 (2007), pp. 107–112.

[225] Athanasios Gazis, Pantelis Karaiskos, and Constantinos Loukas. “Surgical gesture recognition in laparoscopic tasks based on the transformer network and self-supervised learning”. In: *Bioengineering* 9.12 (2022), p. 737.

[226] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *YOLO by Ultralytics*. Available online: `https://github.com/ultralytics/ultralytics` (accessed on 28 February 2023). 2023.

[227] Glenn Jocher. *YOLOv5 by Ultralytics*. Available online: `https://github.com/ultralytics/yolov5` (accessed on 28 February 2023). 2020.

[228] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[229] Kyunghyun Cho. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[230] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[231] Riccardo Franceschini et al. “Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022, pp. 2589–2596.

[232] Lin Yuan et al. “Spatial transformer network with transfer learning for small-scale fine-grained skeleton-based tai chi action recognition”. In: *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*. IEEE. 2022, pp. 1–6.

[233] Amir Shahroudy et al. “Ntu rgb+ d: A large scale dataset for 3d human activity analysis”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1010–1019.

[234] Eddie Bkheet et al. “Using hand pose estimation to automate open surgery training feedback”. In: *International Journal of Computer Assisted Radiology and Surgery* 18.7 (2023), pp. 1279–1285.

[235] Daochang Liu et al. “Diffusion Action Segmentation”. In: *arXiv preprint arXiv:2303.17959* (2023).