Clinical Video Analysis with Geometric Feature Enhanced Deep Learning

XIATIAN ZHANG

Clinical Video Analysis with Geometric Feature Enhanced Deep Learning

Xiatian Zhang

A thesis presented for the degree of
Doctor of Philosophy at Durham University

Department of Computer Science
Durham University
United Kingdom
May 2, 2025

Abstract

Clinical videos are essential in medical intervention, diagnosis, and training, yet their analysis presents substantial challenges due to the complexity and variability inherent in clinical environments. Traditional methods, reliant on manual annotation and human expertise, are limited in scalability and efficiency, particularly in resource-constrained settings. While deep learning offers promising avenues for automation, conventional RGB-based approaches struggle with issues such as occlusions, poor visibility during surgeries, and complex clinical backgrounds. To address these issues, the use of geometric features, such as bounding boxes, depth maps, and human skeleton data, provides a promising solution. These features enable efficient and robust structured understanding in clinical video analysis. This thesis explores how geometric feature enhanced deep learning can address these challenges, focusing on three critical objectives: long-term video anticipation, video quality improvement, and fine-grained semantic understanding.

For long-term video anticipation, a novel adaptive graph learning framework leveraging geometric features as the primary input is proposed for surgical workflow anticipation. This framework introduces a novel geometric representation including bounding boxes of surgical instruments and anatomical targets. Its adaptive graph dynamically selects and updates graph structures to capture the evolving relationships in surgical videos. Validated on two benchmark datasets, this approach demonstrates robust performance across diverse surgical scenarios, offering meaningful predictive insights for surgical teams and semi-autonomous robotic systems.

For video quality improvement, a depth-aware endoscopic video inpainting framework that fuses geometric features and visual features is introduced to address challenges in extreme clinical environments. The framework integrates a Spatial-Temporal Guided Depth Estimation module for direct depth prediction, a Bi-Modal Paired Channel Fusion module for effective visual-depth feature integration, and a Depth-Enhanced Discriminator for assessing the fidelity of reconstructed RGB-D sequences. Unlike traditional 2D-only approaches, this method incorporates depth information, significantly enhancing the realism and spatial accuracy of inpainted content in endoscopic videos.

For fine-grained semantic understanding, multi-view geometric features are integrated into clinical skill assessment frameworks for procedures such as Traditional Chinese Medicine (TCM) physical therapy and Cardiopulmonary Resuscitation (CPR). Two novel publicly accessible multi-view video datasets are introduced for TCM physical therapy and CPR, alongside the Cross-view Multimodality Enhanced Action Quality Assessment framework. This framework combines geometric and visual features for clinical skill assessment, supporting single-view input during inference while retaining multi-view awareness from training. It significantly improves performance in complex tasks such as Needle Depth and Quick Needle Movements. Furthermore, in experiments with the CPR dataset, the proposed framework delivered performance comparable

to that of human experts.

By respectively integrating geometric features as input, for feature fusion, and through multi-view approaches within deep learning frameworks, this thesis demonstrates significant improvements in addressing distinct challenges in clinical video analysis through geometric feature enhanced deep learning. The results hold promising potential for further applications in automated clinical video analysis, including medical intervention, diagnostics, and training. Most of the works have been recognized in peer-reviewed conferences and journals, underscoring their impact and relevance within the field.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Xiatian Zhang is the author's legal name and the transliteration of their Chinese legal name. Additionally, the individual is known by the preferred name Francis Xiatian Zhang, which is used in publications, including works referenced in this thesis, as well as in various professional contexts such as signatures in source code, emails, and documentation.

Acknowledgements

In the PhD journey of my life at Durham, I would like to thank my supervisor, Prof. Hubert P. H. Shum, who has guided me from being a beginner in research to now starting to find joy in research within my beloved field of computer-assisted healthcare.

I would also like to thank my co-supervisor, Dr Noura Al Moubayed, who had a significant impact at the beginning of my research and helped me develop networking skills, which have greatly benefited me.

I extend my gratitude to Dr Jingjing Deng, Dr Robert Lieck, Prof. Xianghua Xie, Dr Merryn Constable, Dr Edmond S. L. Ho, and other invaluable collaborators who have enriched my skills, communication abilities, and cross-disciplinary working experiences, including research, writing, experimentation, and so on. Additionally, I want to express my sincere thanks to my PhD progress reviewers, Prof. Effie Lai-Chong Law and Dr Jingyun Wang, for providing me with consistent and objective feedback, helping me stay on the right path.

I am grateful to my fellow research students at Durham, including Haozheng Zhang, Luca Crosato, Manli Zhu, Mridula Vijendran, Ruochen Li, Ruishen Han, Shuang (Chris) Chen, Tanqiu Qiao, Xiaotang Zhang, Yoshiki Kubotani, and Ziyi Chang in our lab, as well as other students, including but not limited to Li (Luis) Li, Jialin Yu, Mingyue Liu, Mingze Hou, Xingyu Miao, and Zhongtian Sun. They have provided invaluable support during challenging times, helped me improve the quality of my research, and contributed meaningfully to my development.

Additionally, I am grateful for the beautiful natural scenery of the Lake District and Northumberland, which brought me calm and renewal during exhausting times. I also extend my heartfelt thanks to the Jellycat toys my wife and I shared. Their cuteness and comforting presence felt like having a little family member with us through the challenges.

I also want to sincerely thank my parents. Their unconditional support and encouragement gave me the strength to keep going, even in difficult times. Without their understanding and care, I could not have completed this journey.

Finally, I would like to express my deepest gratitude to my wife, Ziyang Yang, who believed in me, stood by my side, and accepted my proposal during the most challenging phase of my PhD. She will forever be my lifelong soul mate.

Dedication

To my family.

Contents

List of Figures

List of Tables

Acronyms

3DGS 3D Gaussian Splatting. 49

AQA Action Quality Assessment. xviii, 37–39, 41, 103, 107, 108, 110–112, 118

AVPF Attention based Visual-Pose Fusion. x, xviii, 102–105, 108, 109, 111, 112, 114, 116, 120

BMPCF Bi-Modal Paired Channel Fusion. xix, 86, 87, 90, 93, 94, 97, 100

CME-AQA Cross-view Multimodality Enhanced Action Quality Assessment. xviii, xix, 14, 17, 18, 102–105, 108, 113, 115, 117, 119–121, 123

CNN Convolutional Neural Network. viii, xiii, 5, 21–25, 30–33, 38, 39, 43, 45, 46

CPR Cardiopulmonary Resuscitation. x, xiv, xviii, 1, 7, 14, 17, 19, 28, 41, 51, 104, 105, 108, 109, 117–119, 121

DAEVI Depth-Aware Endoscopic Video Inpainting. 13, 17, 18, 85–87, 91–93, 96, 97, 99, 100, 123

DED Depth-Enhanced Discriminator. xix, 86, 88, 91, 93, 94, 97, 100

FFN Feedforward Neural Network. 110, 111

FL Federated Learning. 126

FPS Frames Per Second. 72

GAN Generative Adversarial Network. xiv, xix, 35, 91, 93

GCN Graph Convolutional Network. 53

IIA-Net Instrument Interaction Aware Anticipation Network. 52, 74–76, 78, 79

IID Independently and Identically Distributed. 126

LSTM Long Short-Term Memory. 23, 24, 32, 53, 113

MAE Mean Absolute Error. 57, 59, 118–120

MIS Minimally Invasive Surgery. xiii, xiv, 26, 27, 56

MS-TCN Multi-Stage Temporal Convolutional Network. 30

MSE Mean Squared Error. 85, 86, 92, 100

MVA Multiscale View Alignment. x, xviii, 102–105, 108, 109, 111, 112, 114, 116, 121

NeRF Neural Radiance Fields. 49

PSNR Peak Signal-to-Noise Ratio. 85, 86, 92, 100

RAS Robotic-Assisted Surgery. xiv, 9, 26–28, 31, 77

RMSE Root Mean Squared Error. 97

RNN Recurrent Neural Network. viii, xiii, 21, 23, 24, 31, 53

RSD Remaining Surgical Duration. 72, 75, 77, 82

SfM Structure-from-Motion. 48, 49

SIFT Scale-Invariant Feature Transform. 42, 43

SOTA State of the Art. 49, 74, 75, 78, 92, 99

SSIM Structural Similarity Index Measure. 92

ST-GCN Spatial-Temporal Graph Convolutional Network. 53

STGDE Spatial-Temporal Guided Depth Estimation. ix, xvii, xix, 86–89, 93, 94, 96, 97, 100

STTN Spatial-Temporal Transformer Network. 92

SURF Speeded-Up Robust Features. 42, 43

TCM Traditional Chinese Medicine. xviii, 7, 14, 17, 41, 103–106, 108, 113, 119, 121, 125

TCN Temporal Convolutional Network. 30, 67

ViT Vision Transformer. xiii, 24, 25, 44

ViViT Video Vision Transformer. 24

YOLO You Only Look Once. 46, 60, 61, 63

CHAPTER 1

Introduction

Clinical videos are essential resources for clinical intervention [16], diagnosis [17], and training [18]. Traditional analysis methods often require the significant effort of human experts [19], which limits accessibility in developing regions. Recently, the development of deep learning and its application in computer vision has improved this situation by providing low-cost, automatic analysis methods [20]. Deep learning systems process raw RGB clinical videos into latent features and use a learnable approach to infer various predefined tasks such as clinical workflow planning [21], scenario reconstruction [13], and clinical training skill assessment [22]. These advancements significantly enhance the accessibility and convenience of clinical video analysis.

In this thesis, the term *clinical video* mainly encompasses two distinct types: (1) endoscopic videos, which are captured internally during minimally invasive procedures for diagnosis or surgical intervention, and (2) training videos, which are externally recorded using standard RGB cameras during clinical education sessions, such as Cardiopulmonary Resuscitation (CPR) or acupuncture training. These two types of clinical videos are selected because they represent different but complementary challenges for video analysis. Endoscopic videos [23] typically involve constrained viewpoints, variable lighting, and anatomy-instrument interactions, making tasks like future understanding and video

quality enhancement critical. In contrast, training videos feature broader scene contexts with full-body and complex human-object interactions [24, 25], requiring efficient understanding of motion quality. Addressing both types enables a comprehensive exploration of geometric feature integration across varied clinical video scenarios.

While current deep learning methods have improved clinical video analysis, specific environmental conditions in clinical settings continue to pose significant challenges for conventional deep learning frameworks [26]. For instance, intra-operative video feeds during surgeries may suffer from poor visibility due to lighting variations and obstructions such as blood or smoke [27], which often introduce additional noise into the input for deep learning systems, leading to inaccurate analysis outcomes [28]. Additionally, in non-surgical clinical scenarios such as rehabilitation [24], the lack of resilience in conventional deep learning frameworks causes them to struggle with complex hospital backgrounds [29], often preventing them from capturing the most relevant features and delivering accurate inferences. These challenges mean that deep learning systems cannot yet be considered fully reliable for real-world clinical applications [30], necessitating robust solutions that incorporate structured understanding and task-specific prior knowledge to enhance performance [31].

To improve structured understanding and task-specific prior knowledge in deep learning for clinical video analysis, introducing geometric features could be a promising solution [32]. In computer vision, these features capture the shape, spatial relationships, and structural details of objects within video frames. They encompass bounding boxes [21], depth maps [33], and human body skeletons [34], providing structured insights beyond what traditional RGB-based visual features offer. This is especially important in clinical environments where complex backgrounds often obscure critical details [35].

This thesis primarily followed an inductive approach. It was initially motivated by the idea that geometric features could enhance clinical video analysis. The selection of specific features, such as bounding boxes, depth maps, and human skeletons, was progressively refined through experimentation across multiple clinical scenarios. By observing which features most effectively improved performance in tasks such as surgical workflow anticipation, endoscopic video inpainting, and clinical skill assessment, the thesis gradually generalized the role of geometric features as essential components for

enhancing deep learning models in clinical video analysis.

In addition to presenting the research findings, this thesis is structured to serve two primary audiences: researchers entering the field of clinical video analysis and clinicians seeking to understand how advanced deep learning techniques can support and enhance their practice. It aims to provide a complete and independent exploration of how geometric features, including bounding boxes, depth maps, and human skeletons, can be systematically integrated into deep learning frameworks to address real-world clinical challenges. Following a structured progression, the thesis moves from anticipating procedural events to enhancing the quality of clinical imaging and ultimately to assessing clinical skills. Through this unified exploration, it demonstrates the evolving role of geometric feature enhanced deep learning across diverse clinical video scenarios.

1.1 Motivations

1.1.1 Motivations for Clinical Video Analysis

Existing solutions for daily scenario analysis have made notable progress [36], but clinical video analysis presents unique challenges that require dedicated attention. In medical environments, the stakes are considerably higher and errors in video interpretation can directly impact patient safety and treatment outcomes [37]. Complex factors such as variable lighting, occlusions from instruments or bodily fluids, and the dynamic nature of clinical procedures necessitate robust analytical methods that deliver accurate insights in real-time contexts [30]. Furthermore, the varied applications of clinical video [20], which include monitoring surgical workflows, conducting diagnostic assessments, and facilitating clinical training, require tailored solutions that effectively address the specific challenges associated with each context.

There is also a significant need for automated solutions to overcome resource limitations in healthcare [38], particularly in rural regions where expert personnel may be rare [39]. By leveraging advanced automated techniques such as deep learning and incorporating domain-specific insights [38], clinical video analysis could improve accessibility, accuracy, and overall patient care, ultimately bridging the gap between technology and healthcare delivery.

1.1.2 Motivations for Deep Learning in Clinical Video Analysis

Clinical video analysis often requires significant human effort [16]. For instance, in clinical skill assessments, even senior experts must follow complex evaluation procedures that involve repeated video reviews, peer feedback, and facilitator guidance [40]. This reliance on manual interpretation limits the potential of clinical videos, which are easy to capture but challenging to fully utilize, especially in regions lacking sufficient senior, experienced medical staff [41].

Recent advancements in deep learning offer a promising solution to this issue. Deep learning methods learn directly from data, adjusting model parameters through training and feedback from loss functions [1], a process that mirrors how human experts acquire knowledge in medical fields. There is substantial evidence of deep learning's success in medical applications [42], particularly in static medical imaging analysis. For example, the FDA has approved the real-world application of AI-powered screening for diabetic retinopathy using fundus photography [43]. More recently, the application of computer vision in surgical assistance [44] and clinical training evaluation [45] highlights the potential of deep learning in broader clinical video applications beyond medical imaging.

Despite this potential, clinical video analysis poses distinct challenges that differentiate it from traditional deep learning for video analysis, necessitating more advanced solutions tailored to the medical domain. In surgical settings, for instance, visibility issues such as lighting variations [46], smoke [27], or blood [46] make it difficult for conventional methods to extract meaningful information in real time. Similarly, non-surgical clinical environments, such as rehabilitation [24] or emergency care [47], often feature complex backgrounds and dynamic conditions that impede traditional deep learning systems from capturing the most relevant features. Hence, there is still a need for specific designs to introduce deep learning into clinical video analysis.

1.1.3 Motivations for Geometric Feature Enhanced Deep Learning

The introduction of geometric features provides a promising and necessary enhancement to overcome the limitations of traditional deep learning methods based solely on visual features in clinical video analysis [48]. Visual features from RGB frames in common deep

learning methods such as Convolutional Neural Network (CNN) focus on pixel-level information [49], limiting their ability to capture higher-level spatial relationships [50] and making them susceptible to environmental variations in video [51]. Geometric features facilitate a more structured understanding by correlating visual features with structural prior knowledge [48]. These geometric features include bounding boxes [21], depth maps [52], and human body skeletons [22]. They contribute to a more robust understanding of video content [53], which is crucial for analyzing clinical videos in complex and dynamic clinical settings.

There are many different existing geometric features and associated prior knowledge in various clinical applications. In surgical settings, geometric features such as bounding boxes can enhance the monitoring of spatial relationships between tools [21] and anatomy [54] by explicitly locating them, maintaining effectiveness even when traditional visual features are compromised by factors such as smoke or poor lighting [46]. In rehabilitation or emergency care, these features provide clearer descriptions of human motions by isolating and tracking critical body movements within complex environments [22, 55], enabling more accurate and reliable inferences. Therefore, integrating geometric features and their associated prior knowledge into clinical video analysis could significantly improve a system's ability to capture abstract relationships from complex visual information across various clinical applications [56], leading to more accurate and robust deep learning frameworks for clinical video analysis.

1.2 Problem Definitions

This thesis investigates the potential of geometric feature enhanced deep learning in clinical video analysis by addressing three core challenges: long-term video anticipation, video quality enhancement, and fine-grained semantic analysis (*i.e,* evaluating how well a clinical skill is performed). These challenges are uniquely prominent in clinical video analysis and differ significantly from those encountered in general video tasks [57, 58]. Each challenge highlights specific limitations of traditional deep learning models, which are often optimized for general applications and struggle to meet the specialized demands of clinical scenarios. By framing these challenges as key objectives, this thesis aims to

demonstrate how the integration of geometric features can address these limitations and advance the field of clinical video analysis.

1. Long-term Video Anticipation: One of the significant distinctions in clinical video analysis is the prevalence of long-term videos, particularly during procedures such as surgeries, where recordings could span several hours [2]. Unlike recognition tasks in everyday scenarios that primarily focus on understanding the current state, clinical videos require the ability to anticipate future steps to adequately prepare for events before they occur, particularly for tasks like surgical workflow anticipation [4]. This need involves predicting subsequent steps based on long-term temporal dynamics, which is further complicated by the highly variable and dynamic nature of clinical interventions [4]. Standard video analysis methods, which often rely on pixel-level information, struggle to efficiently extract meaningful insights over extended periods, especially when tasked with tracking complex and evolving interactions between instruments and anatomical structures [21]. The detailed definition of this problem is provided in Section 1.2.1.

2. Video Quality Improvement: Clinical videos, particularly those captured in endoscopic procedures, often suffer from poor visibility due to obstructions such as blood, smoke, or fluid, as well as sudden shifts in camera angles within a very narrow field of view [10]. Unlike other domains, clinical settings require not only the recovery of visual data but also the preservation of complex anatomical details [59], which are essential for accurately assessing patients' conditions during examinations. Standard video quality enhancement methods often assume consistent lighting conditions and rely heavily on optical flow as a reference for inpainting flawed regions [60]. Additionally, these methods primarily focus on reconstructing 2D RGB details [61]. These limitations may fail to reconstruct spatial details realistically under variable lighting conditions, as is commonly encountered in clinical video settings. The detailed definition of this problem is provided in Section 1.2.2.

3. Fine-grained Semantic Analysis: *Fine-grained analysis* in video analysis refers to the detailed understanding of *variations within a type of action* [62], meaning the analysis goes deeper than common action recognition. In particular, it involves assessing *how well* an action is executed, which is crucial in clinical settings. In contrast, *coarse-grained analysis* in common video analysis refers to a model's ability to recognize *what type of*

action is being performed by distinguishing between different activity categories [62, 63], which often lacks the capacity for detailed action understanding. Clinical skill assessment requires fine-grained analysis to capture subtle variations in movement and technique, as these can significantly impact a healthcare professional's competency and, ultimately, patient outcomes [64]. For example, recognizing that a surgeon is suturing is not sufficient; it is equally important to evaluate whether the technique is precise and adheres to best practices [65]. Capturing this level of detail is particularly challenging due to common issues such as self-occlusion, where parts of the clinician's body obscure essential actions, and the complex, dynamic environments in which clinical procedures are carried out [66,67]. Standard methods for fine-grained semantic understanding in skill assessment often rely on single-view recordings [68]. As a result, they fail to fully exploit the potential of geometric features for clinical video analysis. Specifically, these methods do not adequately address the need for multi-view understanding [69], which is crucial for effectively capturing the full semantics of movements and interactions in clinical settings. The detailed definition of this problem is provided in Section 1.2.3.

In summary, by focusing on three critical challenges: long-term video anticipation, video quality enhancement, and fine-grained semantic analysis, this thesis demonstrates the power of geometric features in addressing the specific demands of clinical video analysis. These enhancements are crucial for improving real-time processing, increasing robustness against video degradation, and achieving a deeper, structured understanding of medical procedures [35].

While the first two challenges address improvements in endoscopic video analysis, the third challenge extends the investigation to human motion analysis in clinical training scenarios. This extension was motivated not only by practical constraints, such as the limited availability of large-scale annotated surgical video datasets, but also by a scientific need to validate the broader applicability of geometric feature-enhanced deep learning. Both endoscopic imaging and clinical training videos share core challenges, such as occlusions, complex motion patterns, and the need for structured feature extraction [70, 71], but they differ significantly in scene dynamics and task objectives. By addressing clinical skill assessment tasks such as Traditional Chinese Medicine (TCM) physical therapy and Cardiopulmonary Resuscitation (CPR), this research systematically evaluates

whether geometric feature integration can generalize across different types of clinical videos. This broader evaluation framework strengthens the thesis's contribution by demonstrating the versatility and scalability of geometric feature-based approaches beyond inner-body imaging to human-centered clinical activities.

1.2.1 Surgical Workflow Anticipation

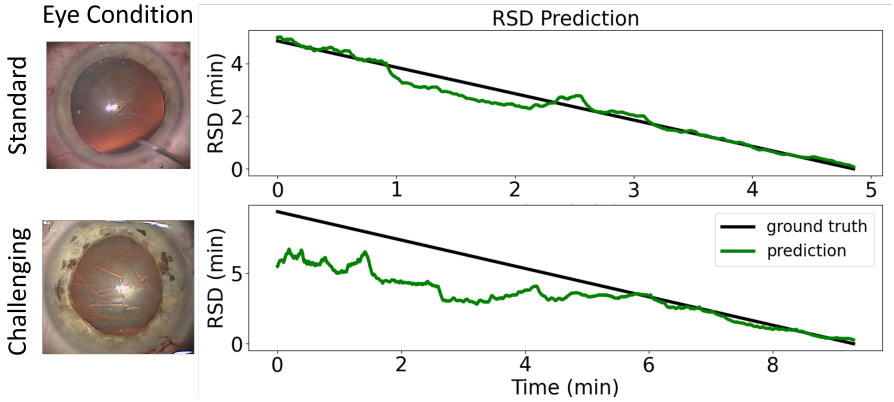

Figure 1.1: An example of common surgical workflow anticipation from Chapter 3, where the end of cataract surgery is treated as an anticipated event. The black line represents the ground truth countdown time to the end of surgery, while the green line depicts the predicted remaining time for surgery, with each prediction at a given time point relying solely on the information available up to that point. The upper part illustrates the anticipation for a standard case with normal eye conditions, while the bottom part represents a more challenging case involving significant eye inflammation. The anticipation is expected to achieve reasonable accuracy across different cases.

Surgical workflow anticipation is chosen over other types of clinical videos to demonstrate the potential of geometric feature enhanced deep learning in addressing long-term video anticipation because surgical videos are uniquely suited to highlight these challenges. These videos are characterized by their extended durations [2], high complexity [21], and the critical need for real-time decision-making [72]. They involve dynamic interactions between surgical instruments, anatomy, and the surgical environment [21], making them an ideal and challenging case study for testing methods that require long-term anticipation. The surgical workflow anticipation task involves predicting future steps from real-time intra-operative video feeds, which is crucial for enhancing surgical decision-making and promoting patient safety [4]. The input consists of the clinical video clip that has already occurred, while the output is the countdown time to the surgical

event of interest, such as switching instruments, initiating different surgical phases, or completing the surgery (see Fig. 1.1).

In detail, this task could be defined mathematically as follows: Given an observed surgical video clip $X_{T_{\text{obs}}} \in \mathbb{R}^{T_{\text{obs}} \times H \times W \times 3}$, where X denotes the RGB frames, T_{obs} denotes the observed time point, and $H \times W$ represents the spatial dimensions, the task is to regress the countdown time $Y_e \in \mathbb{R}^E$ for the surgical event e with a given upper bound horizon h, where E is the total number of events to anticipate. The horizon h specifies that the task only considers the countdown prediction within the h time span. If the regression output exceeds h, the output is reset to h. This adjustment is necessary because entire surgical procedures typically last over 1 hour, and for most specific surgical events, anticipation within a few minutes is more relevant to the current surgical situation [73].

The capability of this task facilitates efficient instrument preparation and surgical phase transitions, enhancing patient safety and fostering smoother communication in the operating room [74]. Furthermore, this approach could enhance the awareness of Robotic-Assisted Surgery (RAS) systems regarding upcoming procedural steps, enabling more timely robotic assistance that could lead to better patient outcomes [75].

1.2.2 Endoscopic Video Inpainting

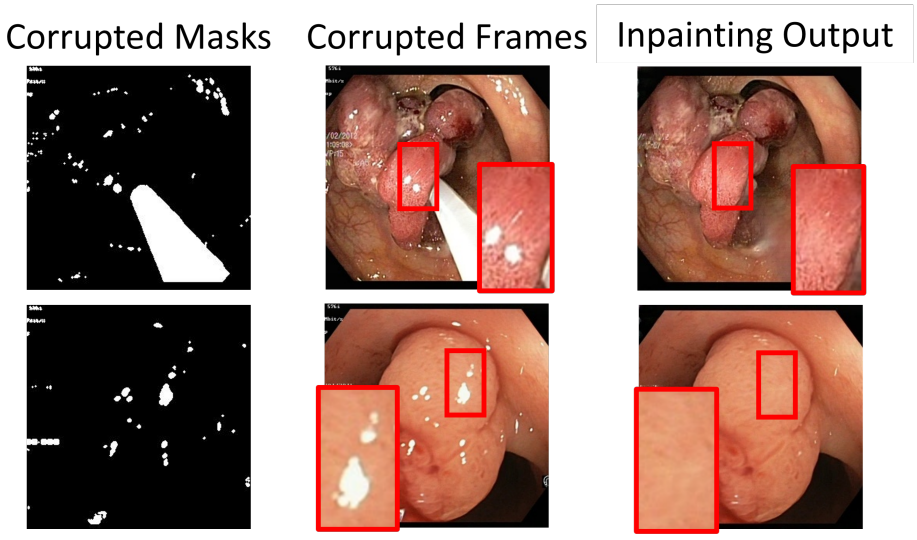

Figure 1.2: An example of endoscopic video inpainting from Chapter 4. The mask indicates the corrupted region to be inpainted, while the red box highlights the area that has been inpainted. In this task, plausible content is expected to fill these corrupted regions.

Endoscopic video inpainting is chosen to demonstrate the potential of geometric feature enhanced deep learning in addressing video quality enhancement because endoscopic videos are inherently prone to obstructions, such as surgical instruments, blood, or smoke, which significantly degrade visual clarity [13]. These obstructions are not only common but also highly disruptive to real-time clinical workflows [10], making them an ideal scenario for evaluating the effectiveness of quality enhancement methods. Unlike general video quality enhancement tasks, which often focus on aesthetic improvements or noise reduction [28], endoscopic video inpainting demands a high level of accuracy to preserve complex anatomical details that are critical for clinical decision-making. The task involves reconstructing missing or obscured video regions to ensure continuous and meaningful video recording, which is vital for real-time surgical monitoring and decision-making systems. The input typically consists of a clinical video clip with corrupted regions, and the output is the same clip with the occluded areas reconstructed to provide plausible details [13] (See Fig. 1.2).

In detail, this task could be defined mathematically as follows: Given the input endoscopic video frames $X \in \mathbb{R}^{T\times H\times W\times 3}$, the framework leverages a binary mask $M \in \mathbb{R}^{T\times H\times W\times 1}$, which identifies the corrupted regions, to produce the modified input $X_M = X \odot M$. Here, $\odot$ denotes the element-wise product. The inpainting framework then processes X_M to generate the uncorrupted output $\hat{Y} \in \mathbb{R}^{T\times H\times W\times 3}$, where $H \times W$ represents the spatial dimensions.

The capability of this task offers a potential solution to enhance the quality of endoscopic videos, facilitating informed clinical decision-making and improving downstream computer vision tasks such as depth estimation [14]. This enhancement could bolster other computer vision-based surgical assistance systems by providing clearer visual information as input [13].

1.2.3 Clinical Skill Assessment

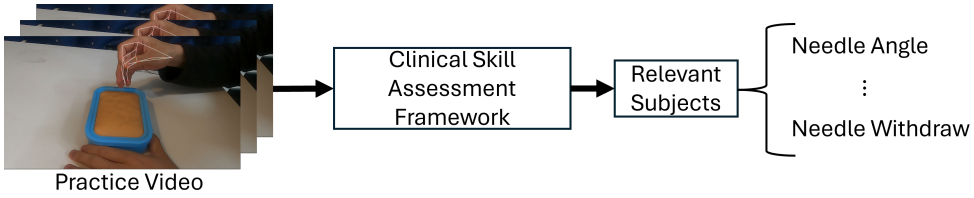

Figure 1.3: An example of clinical skill assessment for acupuncture from Chapter 5. The assessment focuses on the practitioner's hand movements (highlighted by the skeleton of hand joints), and the output includes predictions for different clinically relevant subjects.

Clinical skill assessment is chosen to demonstrate the potential of geometric feature enhanced deep learning in fine-grained semantic understanding for clinical videos, due to its strong reliance on detailed movement analysis in training practices [76]. Unlike general video analysis tasks, which often focus on action recognition [77], clinical skill assessment requires understanding subtle differences in movements within the same training task, making it an ideal scenario for evaluating the effectiveness of fine-grained semantic understanding. This task involves assessing healthcare professionals' proficiency through detailed analysis of their movements and techniques during medical procedures [78]. The input typically consists of raw RGB video clips of a clinical operation, such as acupuncture, while the output is the predicted skill level for clinically relevant parameters observed from the practice movements [68] (see Fig. 1.3).

In detail, this task could be defined mathematically as follows: Given the input clinical practice video, frameworks may take two distinct approaches. Some frameworks directly use the raw RGB video $X \in \mathbb{R}^{T\times H\times W\times 3}$, where $H \times W$ represents the spatial dimensions of each frame, and T is the number of temporal frames. The clinical skill assessment framework is then designed to classify a vector $Y_s \in \mathbb{R}^S$ based on the given video, where s denotes different clinically relevant subjects, and S represents the total number of these clinical subjects.

This task not only promotes the objective evaluation and standardization of clinical procedures in training environments [79] but also has the potential to frequently detect which aspects of clinical training need the most improvement [80]. This enables healthcare professionals to optimize their learning curves more rapidly and maintain high-performance levels [80].

1.3 Research Aims and Objectives

The overall aim of this thesis is to address key challenges in current clinical video analysis through the integration of geometry feature enhanced deep learning. This aim is divided into the following technical objectives, each respectively tackling distinct challenges in long-term video anticipation, video quality enhancement, and fine-grained semantic understanding in clinical video analysis. These objectives employ innovative methods to improve the understanding and processing of video data in medical settings.

1. **Enhance Real-time Long-term Anticipation for Clinical Video:**

    - **Robust Geometric Representation of Surgical Interaction:** To propose a novel geometric feature representation that includes instruments and surgical targets (*i.e.,* anatomy), along with their confidence values, as the primary input of our anticipation framework. The geometric feature representation should provide a robust and structured description of how key surgical elements interact during surgery.
    - **Dynamic Understanding of Surgical Interaction:** To propose an adaptive graph selection method that dynamically selects the optimal graph representation for the interaction relationships between instruments and surgical targets, reflecting the diverse nature of surgical procedures.

2. **Enhance Video Quality Enhancement in Extreme Environments for Clinical Video:**

    - **Efficient Fusion between Visual Features and Geometric Features:** To propose a method that fuses geometric and visual features, allowing geometric features to not only serve as additional inputs but also guide the learning process for enhanced video quality.
    - **3D Geometric Understanding of Endoscopic Scenarios:** To propose a method that incorporates 3D geometric features to enable depth-aware reconstruction in endoscopic scenarios, generating more realistic details in corrupted regions.

3. **Enhance Fine-grained Semantic Understanding in Clinical Training Videos:**

    - **Multi-view Geometric Understanding of Clinical Practice:** To propose a method that incorporates multi-view geometric features for clinical skill assessment, enabling the capture of comprehensive motion details and mitigating the effects of self-occlusion. This approach extends the application of geometric features beyond single-view setups to achieve a more holistic multi-view perspective.
    - **Dataset Creation for Multi-view Clinical Training:** To develop multi-view datasets for clinical training, addressing the current scarcity of public datasets and fostering advancements in clinical video analysis research.

These objectives collectively seek to improve current methodologies in clinical video analysis by integrating geometric features to address challenges that traditional RGB-based methods cannot resolve. By achieving these objectives, this thesis will demonstrate the practical and technical potential of geometric features to advance medical video analysis, leading to improved diagnostics, surgical planning, and clinical training.

1.4 Contributions

The main contributions of this thesis are summarized as follows:

- We propose an adaptive graph learning framework that leverages geometric features as the primary input for surgical workflow anticipation, including both surgical instruments and target anatomy. This framework introduces a novel geometric representation using bounding boxes to extract features of instruments and targets, incorporating their detection confidence levels. Our approach dynamically selects candidate graphs to represent interactions among surgical instruments and targets for each timeframe. By employing graph and temporal convolutions, it effectively utilizes dynamic geometric features, enhancing predictions in complex surgical settings (see Chapter 3).
- We introduce a novel endoscopic video inpainting framework, Depth-Aware Endoscopic Video Inpainting (DAEVI), which efficiently fuses geometric and visual

features. It incorporates depth information to achieve reliable 3D spatial details. This framework extracts depth information during visual feature learning, thus eliminating the need for pre-acquired depth maps as input. It employs a tailor-made feature fusion algorithm to better correlate the 3D spatial relevancy between visual and depth features by pair-wise fusing each visual and depth feature. Additionally, it assesses the 3D spatial fidelity of the RGB-D sequence formed by the inpainted frames and estimated depths, promoting realistic outputs with plausible 3D spatial details (see Chapter 4).

- We introduce the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework, which integrates multi-view geometric and visual features for clinical skill assessment. We also propose two novel publicly accessible multi-view video datasets for Traditional Chinese Medicine (TCM) physical therapy and Cardiopulmonary Resuscitation (CPR). This CME-AQA framework leverages shared-weight cross-attention between views, supports single-view input during inference while retaining multi-view awareness from training, and significantly improves performance in complex tasks such as Needle Depth and Quick Needle Movements. Furthermore, in our experiments with our CPR dataset, CME-AQA achieved performance comparable to human experts (see Chapter 5).

1.5 Publications

The research related to this thesis has been previously published in the following peer-reviewed publications:

- **Zhang, F. X.**, Moubayed, N. A., & Shum, H. P. H. (2022). Towards graph representation learning based surgical workflow anticipation. In *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI '22)*, pages 1-4, IEEE.

- **Zhang, F. X.**, Deng, J., Lieck, R., & Shum, H. P. H. (2025). Adaptive graph learning from spatial information for surgical workflow anticipation. *IEEE Transactions on Medical Robotics and Bionics*, *7*(1), 266–280.

- **Zhang, F. X.**, Chen, S., Xie, X., & Shum, H. P. H. (2024). Depth-aware endoscopic video inpainting. In *Proceedings of the 2024 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '24)*, Springer, Marrakesh, Morocco, pp. 143–153.

- Constable, M. D., **Zhang, F. X.**, Conner, T., Monk, D., Rajsic, J., Ford, C., Park, L. J., Platt, A., Porteous, D., Grierson, L., & Shum, H. P. H. (2024). Advancing Healthcare Practice and Education via Data Sharing: Demonstrating the Utility of Open Data by Training an Artificial Intelligence Model to Assess Cardiopulmonary Resuscitation Skills. In *Advances in Health Sciences Education*, pp. 1-21.

- **Zhang, F. X.**, Yao, H., Chen, S., Jia, X., Zheng, S., & Shum, H. P. H. (2025). Towards cross-view multimodality action quality assessment for Traditional Chinese Medicine physical therapy. Rejected with invitation to resubmit in March 2025; revision in preparation for *IEEE Transactions on Instrumentation and Measurement*.

1.6 Research Chronology

This PhD research began at Durham University in October 2021, with the first two months dedicated to an in-depth literature review on deep learning applications in clinical video analysis and its challenges in surgical video understanding, video quality enhancement, and clinical skill assessment.

In December 2021, the first project on surgical workflow anticipation was initiated, focusing on a fixed graph learning framework integrating bounding boxes for surgical instruments to model their interactions. Early experiments evaluated pixel-based and graph-based approaches, initially using a fixed graph structure for the task. Improvements in workflow anticipation accuracy led to a conference publication in August 2022. Further refinements introduced dynamic graph selection and surgical target integration, significantly enhancing model robustness. This extended work was submitted as a journal paper in November 2023 and later accepted in December 2024.

The second project, started in November 2023, addressed visibility challenges in endoscopic videos through a novel depth-aware inpainting framework. Unlike conventional 2D inpainting, this approach integrated depth estimation and a feature fusion strategy, enabling spatially consistent inpainted frames. It significantly improved anatomical detail preservation, with findings accepted at a conference in June 2024.

The third project originally focused on surgical skill assessment but shifted to multi-view action quality assessment due to ethical constraints and data limitations, broadening the scope beyond surgical videos. To support this, a CPR dataset was collected with Northumbria University (April–September 2022), followed by a multi-view TCM physical therapy dataset (April 2023) in collaboration with Beijing University of Chinese Medicine and Capital Medical University. This led to CME-AQA, a framework integrating multi-view geometric and visual features for clinical skill assessment. The model introduced a cross-view attention mechanism, enabling single-view inference with multi-view awareness. A journal paper based on the CPR dataset was accepted in September 2025. Another paper, submitted in February 2025, received a "revise and resubmit" decision, and the student is currently preparing the resubmission.

The thesis was written from June 2024 to November 2024 and passed with minor corrections in January 2025.

1.7 Thesis Structure

This thesis is organized to systematically explore and present advancements in clinical video analysis by leveraging geometric features for improved accuracy and robustness in medical settings. The chapters are structured to guide the reader through the motivations, literature review, methodologies, and findings of the research in a coherent and logical manner.

Chapter 1: This introductory chapter sets the stage by discussing the motivations behind the research. It highlights the limitations of traditional RGB-based methods in clinical settings and the necessity for advanced geometric feature-based techniques. The chapter outlines the primary challenges in clinical video analysis, such as surgical workflow anticipation, endoscopic video inpainting, and clinical skill rating.

Chapter 2: This literature review chapter evaluates existing methodologies in video analysis, focusing on the transition from RGB-based techniques to the utilization of geometric features. It examines previous works in clinical video analysis, particularly in long-term video modeling, video quality improvement, and fine-grained semantic understanding, establishing a foundation for the novelty of this thesis.

Chapter 3: This chapter introduces our novel geometric feature representation for surgical instruments and anatomy, along with an adaptive graph learning framework for surgical workflow anticipation. It details the methodology and its specific applications in clinical settings, including the use of bounding boxes to extract geometric features of both surgical instruments and target anatomy. These features are integrated into a dynamic selection process for candidate graphs, enhancing prediction accuracy. By leveraging graph and temporal convolutions, the framework effectively interprets dynamic geometric features, addressing the complexities of surgical environments.

Chapter 4: This chapter introduces our novel Depth-Aware Endoscopic Video Inpainting (DAEVI) framework, which efficiently fuses geometric and visual features and incorporates depth information to achieve reliable 3D spatial details. The framework extracts depth information during visual feature learning, eliminating the need for pre-acquired depth maps as input. It employs an efficient feature fusion algorithm to correlate 3D spatial relevance between visual and depth features by pairwise fusing each visual and depth feature. Additionally, it assesses the 3D spatial fidelity of the RGB-D sequence, comprising inpainted frames and estimated depths, ensuring realistic outputs with plausible 3D spatial details.

Chapter 5: This chapter introduces the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework, which integrates multi-view geometric and visual features for clinical skill assessment, along with contributions that include the creation of two multi-view video datasets for TCM and CPR physical therapy. The CME-AQA framework leverages shared-weight cross-attention between views, supports single-view input during inference while retaining multi-view awareness from training, and significantly improves performance in complex tasks such as Needle Depth and Quick Needle Movements. Moreover, in experiments using our CPR dataset, the CME-AQA framework achieved performance comparable to that of human experts.

Chapter 6: The concluding chapter summarizes advancements in the design of adaptive graph methods for surgical workflow anticipation, the development of the DAEVI framework for endoscopic video inpainting, and the creation of two multi-view video datasets alongside the corresponding CME-AQA framework for clinical skill assessment. Additionally, it outlines potential directions for future research, focusing on adaptive geometric feature selection, privacy protection for clinical videos, and other promising avenues.

This structure is designed to offer a clear understanding of the research conducted, from theoretical foundations to practical applications, providing the reader with insights into how geometric features contribute to advanced deep learning in clinical video analysis.

CHAPTER 2

Literature Review

The field of video analysis has undergone significant advancements, driven by the growing complexity of applications across diverse domains, including clinical settings. In healthcare, video analysis is critical in diagnostics, surgical interventions, and skill training, yet its adoption faces unique challenges. These challenges include the need for precise spatial and temporal modeling, robustness against occlusions and artifacts, and adaptability to dynamic environments.

As introduced in Chapter 1, this thesis focuses on two types of clinical videos: endoscopic videos captured during examinations and minimally invasive procedures, and training videos recorded during clinical education sessions, such as CPR or acupuncture. These two types represent complementary challenges for video analysis, requiring solutions that address constrained viewpoints and anatomy-instrument interactions in endoscopy, as well as full-body, human-object interactions in clinical training scenarios.

This chapter systematically reviews prior work along six key dimensions, corresponding to the major research challenges addressed in this thesis. First, it traces the evolution of video analysis methods, from traditional manual and handcrafted approaches (Section 2.1) to modern deep learning frameworks (Section 2.2), establishing the technical foundation for clinical video analysis. Second, it summarizes the specific challenges in

clinical video acquisition and processing (Section 2.3), motivating the need for specialized approaches. Third, it examines techniques for long-term video anticipation (Section 2.4), highlighting their importance in clinical workflows where predicting future actions is critical. Fourth, it explores methods for video quality enhancement (Section 2.5), focusing on their limitations in clinical environments and motivating depth-guided improvements for endoscopic footage. Fifth, it reviews fine-grained semantic understanding approaches (Section 2.6), emphasizing the need for structured, multi-view geometric modeling in clinical skill assessment. Finally, it discusses the role of feature design in video analysis (Section 2.7) and presents how integrating geometric features into deep learning models (Section 2.8) offers a unified pathway to address these challenges across diverse clinical video types.

2.1 Traditional Video Analysis

Traditional methods for video analysis, including manual approaches and early machine learning approaches [81], have long served as foundations in the field. However, these methods come with significant limitations. Manual techniques rely heavily on human expertise to observe, annotate, and interpret video data, making the process labor-intensive and time-consuming [82]. For example, in action recognition tasks, annotators may need to review extensive footage frame by frame to label specific actions [83], such as walking, running, or jumping, which can be particularly demanding for long or complex videos. The dependence on human interpretation introduces the potential for inconsistency, error, and subjectivity, leading to variability in annotation quality and accuracy [84]. These challenges are further amplified when working with large-scale datasets, where maintaining consistency across numerous annotations is difficult.

Early machine learning methods aimed to reduce manual intervention by automating parts of the analysis process. However, these approaches typically rely on handcrafted features, such as color, texture, and motion descriptors [85, 86]. For example, in action recognition tasks, models might use motion descriptors, such as optical flow [87], to detect movement patterns in videos. While this approach could be effective in controlled environments, it often fails in scenes with challenging conditions, such as lighting

variations or reflections [88]. Because these features are manually selected and tuned based on domain knowledge, they lack the flexibility to adapt to diverse or dynamic video content [89]. As a result, traditional machine learning methods struggle to generalize effectively across varied datasets and complex environments.

While manual techniques and early machine learning methods provided a crucial foundation in video analysis, their inherent limitations in scalability, adaptability, and reliability underscore the need for advanced, data-driven approaches [90] that reduce the need for manual input and improve accuracy across a wide range of video analysis applications.

2.2 Deep Learning for Video Analysis

Deep learning has advanced video analysis by enabling models to automatically learn complex patterns directly from raw video data [36]. This approach eliminates the need for manually engineered features and extensive domain-specific expertise. Unlike traditional machine learning methods, which depend on predefined features and manual selection, deep learning models such as Convolutional Neural Network (CNN) [91] and Recurrent Neural Network (RNN) [92] could autonomously extract and interpret meaningful information across spatial and temporal dimensions. By doing so, these models facilitate the automatic extraction of semantic insights from video data, making them highly adaptable to a wide range of applications, from action recognition to object tracking and scene understanding. This capacity for end-to-end learning allows deep learning models to process complex video content efficiently, advancing the automation and accuracy of video analysis workflows.

2.2.1 Convolutional Neural Network (CNN)

CNNs (an example is shown in Fig. 2.1) are among the most widely used deep learning architectures in video analysis. Originally developed for image processing [91], CNNs utilize convolutional layers to automatically extract spatial hierarchies from data. This involves applying various filters that detect low-level features, such as edges and textures, in early layers, and progressively more complex patterns in deeper layers [1]. This hierar-

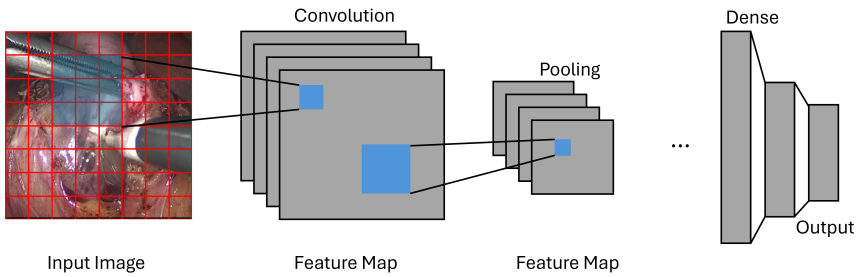

Figure 2.1: Illustration of a typical CNN architecture for video frame analysis. The input frame is divided into a grid, and localized patches are processed through successive convolutional and pooling layers to extract hierarchical spatial features [1]. These features are then flattened and passed through fully connected layers to perform classification or other prediction tasks. In clinical video analysis, such spatial feature extraction is critical for distinguishing clinical contexts, such as stages of intervention [2].

chical feature extraction is particularly beneficial in clinical settings, where distinguishing between complex high-level meanings, such as different interventions, is crucial [93].

The typical framework involves CNNs using raw RGB frames as input, along with corresponding task labels, allowing the deep learning model to learn suitable parameters to predict those labels [35]. For instance, in the case of activity recognition in sports videos [94], a CNN might be used to identify distinct actions, such as running or hitting, by learning spatial patterns associated with each action type across frames. This approach allows the model to automatically extract and interpret key visual features that distinguish each action, improving the accuracy of activity classification in complex and dynamic scenes.

Despite their advantages, relying solely on CNNs for video feature capture could result in the failure to model the relationships between different video frames [95], leading to insufficient modeling of temporal dynamics within the videos. Although there are designs for 3D convolutional networks that enable filters to capture temporal dependencies, the training costs are substantial [95], making a single CNN insufficient for long-term video modeling. Additionally, most CNNs extract features from RGB videos, but the complex clinical settings may lead to varying recording quality [26], causing some extracted features to be unreliable, which limits the effectiveness of CNNs in clinical video analysis.

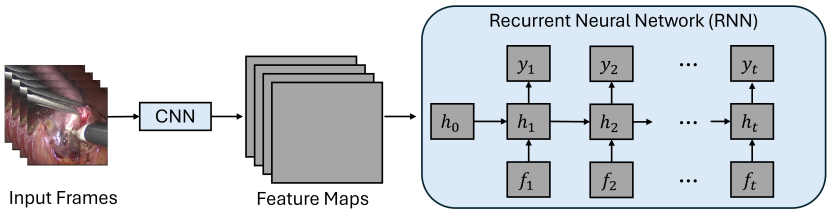

Figure 2.2: Illustration of a hybrid CNN-RNN architecture for video analysis [3]. Individual video frames are first processed by a CNN to extract spatial feature representations $f_1, f_2, \ldots, f_t$. These features are then fed sequentially into a recurrent neural network (RNN), which maintains a hidden state h_t to model temporal dependencies across frames. At each time step, the RNN produces an output y_t based on the current hidden state, enabling sequential predictions. This design effectively captures both spatial and temporal patterns, which is critical for tasks such as clinical activity forecasting [4].

2.2.2 Recurrent Neural Network (RNN)

RNNs are particularly well-suited for analyzing sequential data [96], making them a valuable architecture for video analysis tasks where temporal dependencies are crucial. Unlike CNNs, which focus primarily on spatial hierarchies, RNNs maintain a hidden state that carries information from previous frames, allowing the model to recognize patterns in the temporal flow of events [96]. This ability to process sequences of data makes RNNs ideal for tasks that require understanding the order and progression of frames over time. A classic design for RNN is the Long Short-Term Memory (LSTM) [97], which consists of memory cells and gating mechanisms that regulate the flow of information, enabling the network to learn long-term dependencies while mitigating issues such as vanishing and exploding gradients. This structure allows LSTMs to effectively capture intricate temporal patterns over extended sequences, making them particularly powerful for sequential data analysis.

To enhance the capabilities of both CNNs and RNNs, hybrid models that combine convolutional layers with recurrent architectures (an example is shown in Fig. 2.2) have been developed [96]. These models utilize CNNs to extract spatial features from video frames while employing RNNs to capture temporal relationships between frames. This integrated approach allows for a more comprehensive analysis of video data, effectively addressing both spatial and temporal complexities [35]. By leveraging the strengths of both architectures, these hybrid designs improve the accuracy and robustness of video analysis, enabling better performance in dynamic real-world environments. For example,

Ullah *et al.* [98] leverage LSTM networks with CNN-extracted features to perform action recognition in video sequences with enhanced temporal awareness, achieving better results than methods that lack effective temporal modeling.

Despite the flexibility of these methods, the features learned from RGB frames often fail to capture the complex interactions in videos, such as close human interactions (*e.g.*, pushing, hugging, and high-fiving) [3], leading to potential inefficiencies in modeling and prediction. In addition, the strong assumption in RNN regarding the relationships between adjacent frames differs somewhat from real clinical videos [75], where earlier processes may influence subsequent steps that are far removed from the current time point, rather than just those immediately preceding it. As a result, the RNN may fail to capture the abstract and long-term global interactions between clinical steps [75].

2.2.3 Transformer

Transformers have emerged as a powerful alternative for modeling sequential data, particularly in video analysis, due to their ability to capture long-range dependencies and contextual relationships without the limitations associated with recurrent architectures [99]. Unlike traditional RNNs, which process sequences in a stepwise manner, transformers employ a self-attention mechanism that allows them to weigh the significance of different parts of the input sequence simultaneously. This characteristic enables transformers to learn complex temporal patterns across all frames in a video [100], rather than just those in immediate succession.

In video analysis, transformers enhance the interpretation of sequential data by effectively incorporating both spatial and temporal dimensions [101]. Generally, two design strategies are common: some architectures first extract frame-level features using a CNN and then feed these features into a transformer for temporal modeling [100], while others directly process raw frames without any convolutional operations. For instance, pure transformer-based models, such as the Vision Transformer (ViT) [5] (example ViT design shown in Fig. 2.3), operate differently: they divide input frames into fixed-size patches, linearly embed these patches, add positional encodings, and input them directly into transformer layers without using any convolutional neural networks. In addition, Arnab *et al.* [102] propose Video Vision Transformer (ViViT), which extends the ViT

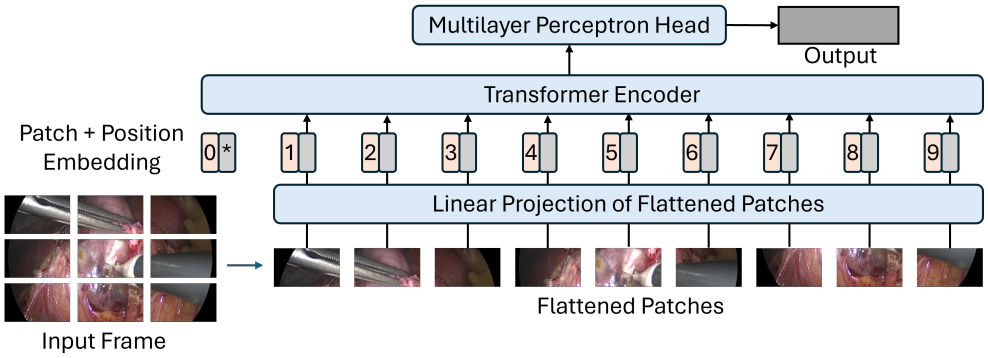

Figure 2.3: Illustration of a Vision Transformer (ViT) architecture for video frame analysis [5]. Each video frame is divided into fixed-size patches, which are flattened and linearly projected into patch embeddings. A special classification token (CLS token, labeled as 0*) is prepended to the sequence. Positional embeddings are added to preserve spatial information. The sequence is then processed by a transformer encoder composed of multi-head self-attention layers, allowing the model to capture spatial and contextual relationships among patches. Finally, the output corresponding to the classification token is fed into a multilayer perceptron head for prediction tasks. This design enables learning global spatial patterns across frames without relying on convolutional operations, which is critical for complex scene understanding in clinical video analysis.

architecture to videos by applying factorized spatial and temporal attention mechanisms on the patch embeddings. This design captures spatial-temporal relationships across the entire sequence without relying on CNNs.

Despite their advantages, transformers require substantial computational resources, particularly for long video sequences, as their self-attention mechanism scales quadratically with input length [103]. Therefore, in very long videos, such as surgical videos that often exceed 30 minutes [2], the efficiency of transformers becomes a significant challenge. Additionally, training vision transformers to effectively grasp a structured understanding for fine-grained semantics, such as the skill level of a clinical operation in training video [104], necessitates a large dataset to develop a reasonable attention matrix, which is often impractical in clinical settings where dataset sizes are limited [105].

In summary, while deep learning methods show significant promise for video analysis, challenges remain even in advanced transformer methods, particularly regarding the computational cost associated with long videos, sensitivity to recording quality, difficulties in achieving a structured understanding of complex interactions, and the challenge of developing fine-grained semantic understanding. To overcome these challenges at the model level, this thesis proposes enhancing deep learning models with geometric features,

such as object locations, depth information, and human skeleton data. By incorporating these geometric cues, the proposed framework aims to reduce computational cost through more efficient representations (Chapter 3), improve robustness against visual artifacts (Chapter 4), and enable a more structured understanding of complex interactions and fine-grained semantics in clinical videos (Chapter 5).

2.3 Clinical Video Acquisition and Challenges

Video analysis in clinical settings is essential for a wide range of applications, including surgical workflow understanding [35], skill assessment, and patient monitoring [63]. As a result, clinical videos could be acquired at various stages of medical procedures, including intervention, diagnosis, and training. Each of these stages presents unique challenges for video analysis.

2.3.1 Future Understanding in Intervention

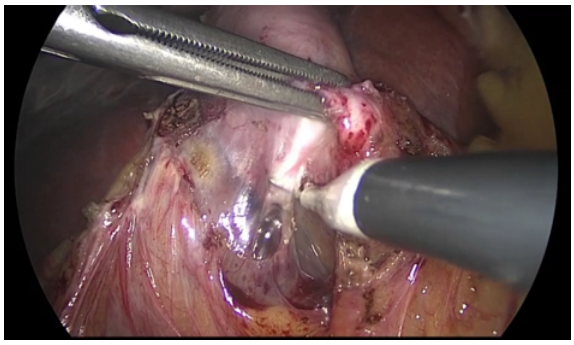

Figure 2.4: An example frame of Minimally Invasive Surgery (MIS) from the Cholec80 dataset (CC BY 4.0) [2]. The setup involves inserting instruments through small incisions in the patient's body, allowing clinicians to operate with minimal physical intrusion. Real-time surgical video is captured during the procedure, providing high-precision visual feedback and enabling effective manipulation of surgical instruments.

In intervention settings, clinical video acquisition is most common in surgical environments, particularly in Minimally Invasive Surgery (MIS) and Robotic-Assisted Surgery (RAS) [106] (see Fig. 2.4). In these scenarios, clinicians make critical, real-time decisions by closely observing video streams captured through *endoscopic cameras*. These small, single-view cameras are mounted on a holder or robotic arm and are inserted into the

patient's body through a small incision, providing internal views of the surgical field [107]. The camera transmits high-definition video to a monitor in real-time [108], enabling the surgeon to navigate and perform procedures with minimal physical intrusion. This setup provides continuous visual feedback, helping the surgeon make precise, anatomy-based decisions.

Understanding the future course of action in surgical video analysis is particularly important to ensure smooth communication between the surgical team and safe interventions [21]. Predicting upcoming surgical phases or anticipating necessary adjustments enables the surgical team to prepare the appropriate tools and make timely decisions [21], ultimately improving workflow efficiency and patient outcomes. In longer surgeries, this forward-looking analysis is crucial for reducing procedural delays and preventing complications that may arise from unexpected developments [4]. However, achieving accurate future understanding presents significant challenges. The narrow field of view [10], variable lighting conditions [109], and potential obstructions, such as smoke or blood [109], complicate real-time video analysis. These factors, combined with the inherent complexity and variability of surgical procedures, make it difficult for current video analysis methods to reliably predict the sequence of actions or identify the next steps in dynamic surgical environments.

2.3.2 Video Quality in Diagnosis

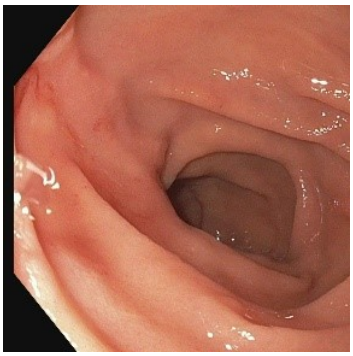

Figure 2.5: An example frame of gastrointestinal endoscopy for inner-body examination from the HyperKvasir dataset (CC BY 4.0) [6]. Endoscopy is a flexible and minimally invasive diagnostic procedure compared to MIS and RAS, as it allows real-time video to be captured by inserting the endoscope through natural openings, such as the mouth, without the need for surgical incisions.

For diagnosis, endoscopic video is commonly acquired in inner-body examinations (see Fig. 2.5), particularly in areas such as the gastrointestinal tract [110] and respiratory system [111]. The camera is often inserted through a natural orifice, such as oral or nasal, rather than creating a new incision as in RAS [112], providing vital visual information that helps clinicians identify abnormalities, such as polyps [113], tumors [114], or signs of infection within internal organs [115].

Clear and high-quality video is essential in endoscopic diagnostics [13], where accurate visual assessment directly impacts early diagnosis and disease management. Improved video quality allows clinicians to detect and analyze potential pathologies more effectively, enhancing the overall reliability of the diagnostic process. Nevertheless, endoscopic video analysis faces significant challenges. Obstructions, such as mucus or blood [116], and variations in lighting conditions [14] can often interfere with real-time analysis, complicating efforts to obtain consistent and clear visuals needed for precise diagnosis.

2.3.3 Fine-grained Semantic Understanding in Training

Figure 2.6: An example of a training video recording for CPR from the dataset we collected in Chapter 5. The recording setup for these videos is simpler compared to endoscopic videos. In this case, a subject performs a practice captured from multiple cameras.

In training contexts, third-person videos captured by standard RGB cameras are commonly used (see Fig. 2.6), particularly in areas such as physical therapy [117] and nursing skills development [118]. These videos provide a comprehensive view of participants' actions, allowing trainers to evaluate body movements, posture, and interactions with equipment or simulated stations, such as dummy pads. This fine-grained semantic understanding is essential in assessing how well practitioners adhere to training guide-

lines and standards during simulated scenarios [119]. Trainers also rely on contextual information, such as participants' interactions with their environment, to deliver accurate skill assessment and feedback [120].

Achieving this level of detailed understanding in video analysis is challenging. Factors such as self-occlusion, where parts of the body obscure one another [121], and complex recording backgrounds, often found in hospital training settings [122], could hinder the accurate analysis of actions. These challenges are further complicated when single-view recordings are used [15], as they limit the depth and perspective needed to fully capture nuanced interactions and movements essential for effective training feedback.

In summary, clinical videos acquired during interventions, diagnoses, and training present unique challenges for video analysis, including long-term future understanding under narrow fields of view and obstructions, maintaining consistent video quality despite dynamic artifacts like smoke and highlights, and achieving fine-grained semantic understanding despite occlusions and limited viewpoints. To address these clinical data challenges, this thesis leverages geometric feature enhanced deep learning, including structured spatial modeling for surgical workflow anticipation (Chapter 3), depth-guided video inpainting for robust endoscopic video quality improvement (Chapter 4), and multi-view skeleton modeling for comprehensive clinical skill assessment (Chapter 5).

2.4 Anticipation in Long-term Video Analysis

Future understanding is essential in clinical video analysis, but the complexity and length of clinical videos present challenges not typically encountered in everyday video scenarios. While many existing solutions focus on long-term future understanding through action anticipation in general contexts [123], the unique demands of clinical videos, particularly in surgical settings [2], require specialized approaches. These methods must address the extended duration and evolving contexts within clinical environments, such as frequent interactions between instruments and anatomy [21], as well as unexpected events like bleeding [124]. Despite these challenges, real-time understanding remains crucial for ensuring timely decisions [35], further highlighting the difficulty of applying such methods effectively in clinical settings.

2.4.1 Long-term Video Understanding

Transformers have emerged as a pivotal architecture in the analysis of long video sequences due to their ability to capture complex temporal relationships. To enable efficient long-sequence modeling, a common approach is to perform self-attention between frames only after extracting visual features using an existing CNN model. Self-attention allows the model to focus on different parts of the input sequence, enhancing its capability to capture long-term dependencies. For example, Zhou *et al.* [125] utilize a multi-layer CNN to extract spatial features from videos, which are then fed into a transformer model to capture long-term relationships between frames for accurate video captioning that describes what happens in a video. However, these models require substantial computational resources [103], as the self-attention mechanism scales quadratically with the number of frames, making them computationally intensive for long sequences.

To achieve a more computationally efficient solution for long-term video understanding, Temporal Convolutional Network (TCN) [126] has been proposed. TCNs effectively handle longer sequences by leveraging dilated convolutions to expand the receptive field without significantly sacrificing performance. This enables TCNs to capture long-term relationships while maintaining a reasonable computational cost. For example, Farha *et al.* [127] propose Multi-Stage Temporal Convolutional Network (MS-TCN), which connects multiple TCN layers in series to learn 3D CNN features for each small clip, allowing for more refined temporal feature learning in action segmentation tasks. The advantages of low-cost online inference and ease of training make TCNs a popular alternative for long video analysis [35], even though TCNs may be less powerful compared to transformers.

Despite these advancements, existing solutions provide limited understanding for long-term video analysis, as most focus only on recognizing the current situation [35], such as clinical intervention recognition. However, in many assistance systems for clinical video, it is essential not only to understand what is happening now but also to predict future events. This capability is crucial for keeping the clinical team informed about upcoming steps, significantly enhancing the potential for real-world applications while improving intervention safety [21].

2.4.2 Long-term Video Action Anticipation

In general computer vision, action anticipation is a common task aimed at achieving a better understanding of future states in video sequences [81]. This capability is particularly valuable in clinical video analysis, as action anticipation helps predict the next steps in surgical procedures or other clinical interventions [21], thereby enhancing decision-making and improving patient outcomes.

The common solution for long-term action anticipation is to predict a sequence of future actions and their durations by capturing historical action features [81]. This is accomplished by dividing the video into smaller, equally sized segments called *snippets* [81], each representing a short time interval. These snippets are then processed using pre-trained models, such as CNNs, to extract meaningful spatial and temporal features. Additionally, RNNs or transformers are employed to analyze the relationships and dependencies among these snippets over time, enabling effective forecasting of future actions. For example, TempAgg [128] extract frame-level features from visual modalities using pre-trained networks like I3D, pool them into multi-scale snippet features via max-pooling, and apply self-attention through non-local blocks to couple recent and long-range temporal representations for action anticipation.

Though these existing solutions for long-term action anticipation demonstrate reasonable performance on benchmarks for everyday scenarios [81], such as cooking [129], where the procedure is quite easy to understand, the differences between everyday scenarios and complex, dynamic clinical video recordings may necessitate a specific design for anticipation applications in clinical settings to provide a robust understanding of complex clinical procedures and varied recording scenarios.

2.4.3 Surgical Workflow Anticipation

A current popular adaption for action anticipation into the clinical setting is surgical workflow anticipation. Surgical workflow anticipation supports RAS by predicting the next likely step a surgeon will take [21]. Unlike imminent tasks such as movement prediction [130–132], this approach focuses on long-term event anticipation. Currently, most methods rely on pixel-level visual features extracted using established visual feature

extraction frameworks [4, 21, 73, 133–135]. Among them, early work such as TimeLSTM [136] combines CNN-extracted features with LSTM networks for surgical phase anticipation, requiring phase annotations during training. RSDNet [133] improves its usability by directly predicting the remaining surgery duration without relying on phase labels.

Though specialized solutions exist to make these methods suitable for surgical scenarios, they often introduce auxiliary tasks, such as surgeon experience classification [134], to enhance the model's understanding of complex surgical semantics. In addition, some methods incorporate a specially designed dropout strategy [4] in their network design to increase robustness in complex surgical scenarios. However, as they only use extracted RGB visual features to represent the spatial features of surgical scenarios, they often fail to capture important structured and abstract relationships within surgical videos, such as instrument interactions [21]. As a result, these limitations significantly affect their anticipation performance due to an insufficient understanding of surgical progress.

In summary, long-term surgical workflow anticipation is a meaningful task that addresses the deficiencies in feature understanding in current long-term clinical video analysis. However, current solutions often miss the structured understanding of complex surgical interactions. To overcome these limitations, this thesis proposes integrating geometric features, such as object locations and structured interaction modeling, into the anticipation framework. By combining geometric information with deep learning, the proposed approach aims to improve the understanding of surgical workflows and enhance long-term prediction accuracy in complex clinical scenarios (Chapter 3).

2.5 Video Quality Improvement

Video quality is a critical factor influencing the complexity of clinical diagnosis analysis and the accuracy of downstream computer-assisted systems [13]. While existing methods in general computer vision provide solutions for video quality improvement [28], clinical videos present unique challenges. A major issue is the lack of ground truth [13], which complicates the optimization of deep learning models. Moreover, clinical videos require not only 2D texture fidelity, which many existing video quality improvement methods

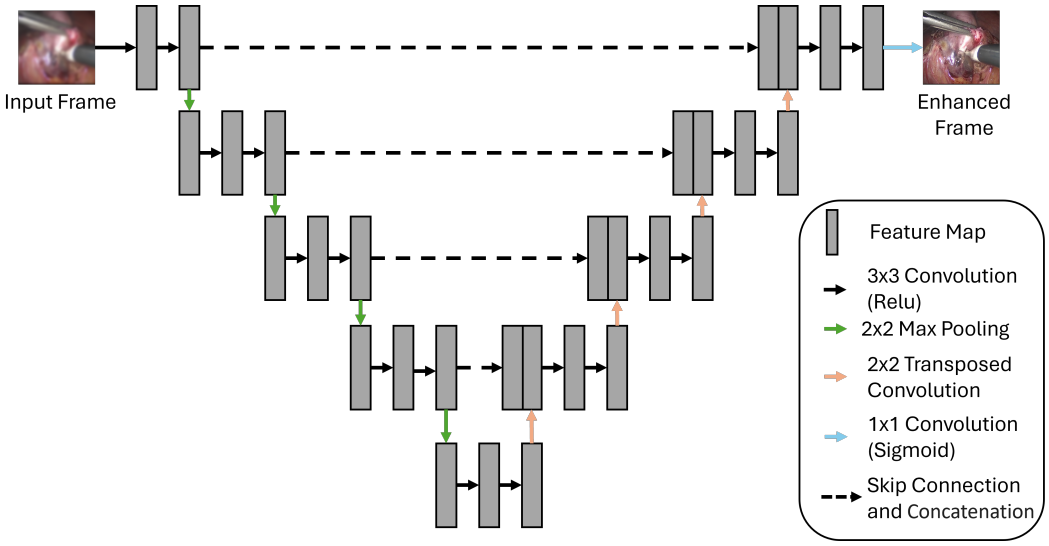

Figure 2.7: Illustration of the U-Net architecture for frame-level enhancement [7, 8]. The input frame passes through a contracting path (left side) composed of repeated convolutional and downsampling (green arrows) operations, capturing contextual features at multiple scales. The expanding path (right side) uses upsampling (orange arrows) and concatenation with corresponding features from the contracting path via skip connections, enabling precise localization and reconstruction of fine details. The final output is a high-quality enhanced frame. This encoder-decoder structure with skip connections helps preserve local spatial details, which is essential for improving clinical video frames affected by noise, occlusion, or low visibility.

aim to optimize [28], but also 3D fidelity to help reconstructed or enhanced content better reflect its meaningful spatial structure for downstream tasks [137]. These challenges underscore the need for specialized designs tailored to clinical video quality enhancement.

2.5.1 Video Quality Enhancement

Deep learning has emerged as a powerful approach for enhancing video quality by leveraging complex patterns and features from the data, where a common solution is an image enhancement, defined as the process of improving the visual quality of an image to make it more suitable for analysis or interpretation [28]. One classic model for improving frame quality in video is the U-Net architecture [7] (example U-Net design shown in Fig. 2.7). U-Net is a CNN originally designed for biomedical image segmentation. Its architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. This structure is characterized by skip connections that link the encoder and decoder layers, allowing the network to retain high-resolution

features that can be beneficial for detailed image processing. In the context of video quality enhancement, for example, Huang *et al.* [8] utilize a U-Net with added average pooling and multiplication between the encoding and decoding features. This approach effectively captures the overall characteristics of the input image, significantly improving its perceptual quality on mobile devices by alleviating visual artifacts.

While there is significant potential for enhancement, these previous deep learning solutions often struggle to handle the unique challenges presented by clinical video data [10], especially in dynamic environments where occlusions and artifacts are common [10]. Common enhancement methods primarily focus on adjusting image quality across the entire frame [138], which could ignore localized artifacts that are more prevalent in clinical videos [139]. Furthermore, classic enhancement techniques tend to concentrate on improving individual frames without adequately considering the temporal context between video frames [28]. This failure could lead to inconsistencies and suboptimal results, as the lack of utilization of the temporal information inherent in video sequences may result in a less cohesive visual experience, ultimately impacting the analysis performed by computer-assisted systems.

2.5.2 Video Inpainting

Another critical method for improving clinical video quality is video inpainting, which specifically targets reconstructing missing or corrupted regions within video frames [140]. Video inpainting goes beyond general enhancement by focusing on restoring visual continuity in localized areas affected by corruption. In this process, neighboring frames serve as references to generate plausible content in the corrupted regions [61], enabling effective interpolation of information. This approach is particularly beneficial in clinical settings, where occlusions can obscure vital anatomical details or affect downstream computer vision assistance systems [13]. Classical methods, such as the automatic region-filling approach proposed by Arnold *et al.* [141], which uses motion-compensated interpolation and spatio-temporal consistency to guide the filling process, and the patch-based optimization framework introduced by Newson *et al.* [12], which globally matches and aggregates similar spatio-temporal patches, demonstrate promising results. However, these methods assume relatively simple motion patterns and either static or smoothly

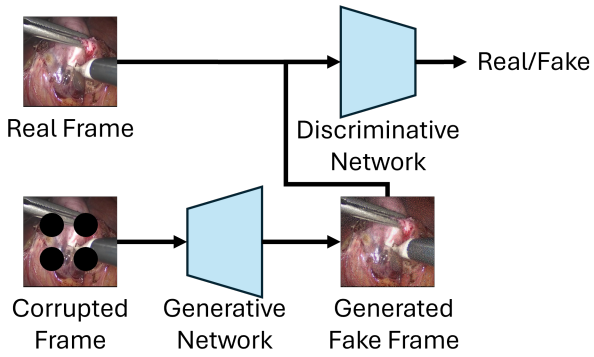

Figure 2.8: Illustration of a GAN-based video inpainting framework [9]. The input consists of a corrupted video frame with missing regions (black holes). A generator network synthesizes the missing content by leveraging information from the surrounding context and temporal clues. The output is a fully reconstructed frame where corrupted areas are realistically restored. A discriminator network evaluates the authenticity of the generated frame, distinguishing between real (ground truth) and fake (generated) frames. This adversarial training setup encourages the generator to produce visually coherent restorations, which is crucial for clinical video applications where preserving plausible and fine-grained details is essential, yet often hindered by occlusions [10].

varying backgrounds, which limits their applicability in complex scenarios.

Recent advancements in deep learning have significantly improved video inpainting approaches, enabling networks to learn from extensive training data and generate more diverse content. To train efficiently, Generative Adversarial Network (GAN) [142] (example GAN-based inpainting shown in Fig. 2.8) is a common training strategy to enhance the quality of inpainting results [140], as it could effectively learn to generate realistic and coherent visual content by minimizing the difference between generated frames and ground truth data. This adversarial training involves two networks: a generator that creates inpainted frames and a discriminator that evaluates their authenticity, thus driving the generator to produce more accurate and high-fidelity reconstructions. For example, Zeng *et al.* [61] use temporal patch-GAN to enhance their video transformer, where they split the input video clip into smaller patches and compute the correlation between these patches to capture spatial-temporal correlation. During their training, temporal patch-GAN uses a 3D convolution discriminator to assess the generated content in the temporal dimension, resulting in outputs that are more temporally consistent when optimizing with these adversarial losses from discrimination.

With these advancements, transferring video inpainting techniques into clinical video

still presents challenges. One challenge is the online inference ability [10]. Most video inpainting methods require dense reference points from both past and future frames to make inferences [60, 61, 143]. However, in some clinical scenarios, such as endoscopy checks, a fast and timely online assistance may be necessary. Another challenge is the lack of ground truth in real clinical videos [13]. In real-world recordings, corrupted regions, such as occlusions or specular highlights, occur naturally during video capture rather than being artificially introduced [13]. As a result, it is impossible to obtain corresponding uncorrupted reference frames for these regions. This differs from common video inpainting benchmarks, where corruption is artificially introduced into clean videos, and the ground truth is readily available for supervised learning [144].

2.5.3 Endoscopic Video Inpainting

To tackle the specific challenges in clinical videos, several attempts have been made, especially for endoscopic video inpainting [10, 13, 46, 145]. Improving the quality of endoscopic videos is essential for enhancing diagnostic accuracy and facilitating downstream computer vision tasks by reducing noise. High-quality footage allows healthcare professionals to observe critical anatomical structures and conditions more effectively [10], leading to clearer observations of inner-body situations. Additionally, clearer videos benefit automated systems reliant on computer vision techniques, as they can operate more effectively by minimizing noise and artifacts [13]. This results in more reliable outputs for tasks such as key region tracking or 3D reconstruction [13], ultimately supporting robust automated analysis in clinical environments.

To facilitate the potential of online video inpainting, Tukra *et al.* [10] propose STV-Net, a densely connected fully convolutional encoder-decoder network that includes 2D and 3D convolutions to capture both spatial and temporal relationships within videos. They train the network with a bespoke loss function combining adversarial, reconstruction, perceptual, style, and temporal losses to achieve accurate and temporally coherent occlusion reconstructions on endoscopic video data. Their methods have been tested on both endoscopic view video and natural scenes, achieving reasonable performance.

To overcome the lack of ground truth, Daher *et al.* [13] propose a pseudo mask generation method for highlighting inpainting in endoscopic videos. They first detect

the corrupted highlight regions in the raw video and then shift the corrupted mask to the right, translating the detected specular masks to new positions over non-occluded regions, ensuring that the areas covered contain known textures. This process allows the model to be trained in a fully unsupervised manner by providing paired data for training, thus enabling effective inpainting of occluded areas caused by specular highlights in endoscopic video streams. Based on this pseudo ground truth, they train a spatial-temporal transformer [61] and find that their inpainting model could achieve reasonable performance and help downstream tasks such as point matching.

While these advancements are noteworthy, current methods [10, 13] often rely on inpainting frameworks based on 2D RGB or optical flow information. In clinical scenarios, such as endoscopic video, the relationships between RGB and spatial information are not easily matched, as the boundaries of different anatomical structures in endoscopic views are often vague [146]. This leads to challenges in preserving vital 3D spatial details, resulting in artifacts and spatial inconsistencies in the inpainted regions, which limits their reliability for clinical applications.

In summary, video inpainting could be an important task for improving clinical video quality. However, the current clinical applications of video inpainting [10, 13] may not sufficiently account for the complex environments in clinical settings, limiting their ability to generate realistic content. To address these limitations, this thesis proposes a depth-guided endoscopic video inpainting framework that leverages geometric features to better preserve spatial consistency and anatomical structures during inpainting. This approach aims to enhance the reliability of video quality improvement in clinical environments, as detailed in Chapter 4.

2.6 Fine-Grained Semantic Analysis in Video

Fine-grained semantic analysis is crucial for clinical videos, as it helps clinicians evaluate whether a clinical operation has been performed correctly. While general approaches, such as those used in Action Quality Assessment (AQA) for sports or daily activities [68], have been successful, they often fail to meet the specific demands of clinical training. Clinical scenarios require evaluating complex details, such as hand motions and human-

object interactions [147], which are critical for tasks like emergency nursing skill training [148] or physical therapy assessment [149]. These requirements highlight the need for tailored solutions that address the unique challenges of clinical video analysis.

2.6.1 Action Quality Assessment

In general computer vision, Action Quality Assessment (AQA) is a task that requires a fine-grained understanding of videos to evaluate the quality of actions performed within them [68, 150]. The common pipeline for AQA typically begins with feature extraction, where deep learning methods are employed to obtain relevant features from the video frames. One of the most commonly utilized architectures for this purpose is the 3D CNN [68]. This architecture uses 3D convolutional layers to capture both spatial and temporal information from video sequences, enabling a comprehensive representation of features. Following feature extraction, the extracted features are processed through an action evaluation module, which typically employs one of two evaluation approaches: regression scoring [151, 152], which predicts a continuous score based on the extracted features and is often used in sports contexts, and pairwise sorting [153, 154], which assesses the quality of actions by comparing pairs of videos to determine which action is of higher quality based on the features extracted.

While these frameworks have shown reasonable performance in distinguishing fine-grained differences in the same action across different subjects, their reliance on single-view analysis [68] limits their applicability to clinical settings. In AQA for clinical videos, the focus often shifts to evaluating dense and complex hand motions or interactions with tools [147], which require capturing detailed spatial and temporal relationships. Single-view methods may miss critical context information, making it difficult to accurately assess complex actions. Therefore, clinical applications necessitate multi-view perspectives or enhanced spatial modeling to capture the nuanced details required for reliable performance assessment.

2.6.2 Clinical Skill Assessment

Clinical skill assessment represents a specific application of AQA in clinical settings, aimed at evaluating the proficiency of practitioners in performing clinical tasks through video analysis. This approach typically focuses on grading actions into skill levels, such as junior or expert, based on a range of performance aspects [80]. Most existing clinical skill assessment studies have focused on surgical skill assessment [80] and rehabilitation assessment [155]. The methodologies for surgical skill assessments are largely similar to general AQA pipelines in computer vision, beginning with feature extraction from clinical videos and followed by evaluation modeling to generate a skill score. These studies frequently employ CNN backbones to extract latent features from video clips [156, 157], which are then used for further modeling and assessment. In contrast, rehabilitation assessments often rely on keypoint extraction algorithms to model joint movements and body interactions [158, 159].

Although these works have demonstrated significant performance in their respective domains, a notable challenge persists with general AQA methods: the reliance on single-view analysis. This limitation reduces their effectiveness in clinical practice, where tasks often involve intricate hand interactions and detailed tool usage [147]. In real-world training and evaluations, critical details may be obscured by self-occlusion or the complexity of interactions with medical tools [69]. While Abdelaal *et al.* [69] proposed a multi-view system for surgical skill assessment, their approach relies on specially designed feedback devices for tracking surgical tool movements, resulting in a highly specific assessment method that is far from a generalized solution. Therefore, incorporating multi-view perspectives remains an ongoing challenge in clinical skill assessment, requiring further development to provide more comprehensive and scalable practice evaluation methods [160].

2.6.3 Current Datasets for Clinical Skill Assessment

One reason current clinical skill assessments often ignore the importance of multi-view analysis is that most publicly available datasets involve only single-view recordings [168]. A summary of these existing public datasets could be found in Table 2.1. From this

Table 2.1: Summary of Publicly Available Datasets for Clinical Skill Assessment

Dataset	Video Number	Subject	Modality	View	Skill Assessment Related Annotation
JIGSAWS [161]	103	8	RGB + Kinematics	1	Surgical Skill Assessment
Cataract-101 [162]	101	101	RGB	1	Surgical Experience Classification
SimSurgSkill2021 [163]	315	Simulation	RGB + Bounding box	1	Surgical Skill Assessment
Hei-Chole [164]	33	33	RGB	1	Surgical Skill Assessment
UI-PRMD [165]	1000	10	RGB + Kinematics	1	Rehabilitation Assessment
UpSurgeon [166]	15	15	RGB	1	Surgical Experience Classification
Keraal [167]	2622	31	RGB + Kinematics	1	Rehabilitation Assessment

table, we can see that the current publicly available datasets primarily use single-view recordings, which limits the depth of analysis for deep learning frameworks in clinical skill assessment. Therefore, it may be necessary to introduce datasets with more diverse view captures.

In summary, AQA provides a potential solution for understanding the clinical video semantics in a fine-grained way, particularly in the context of clinical skill assessment. However, current clinical skill assessments often rely on single-view recordings, which may be affected by frequent self-occlusion in clinical practice, thereby impacting their assessment accuracy. To address these limitations, this thesis proposes a multi-view geometric feature-enhanced framework that models human skeleton interactions across multiple viewpoints, aiming to improve fine-grained semantic understanding and enhance the robustness of clinical skill assessment. Furthermore, to support research in this area, we introduce two new multi-view clinical skill assessment datasets for TCM physical therapy and CPR, specifically designed to capture complex interactions and mitigate the limitations of existing single-view datasets. These contributions are detailed in Chapter 5.

2.7 Features in Video Analysis

Features are essential components of video analysis, providing the foundation for analytical frameworks in both general and clinical contexts. Selecting the appropriate features is crucial for achieving accurate and efficient analysis, particularly in clinical settings where precision and reliability are paramount. In video frames, features can be broadly categorized into visual features and geometric features [48], each offering unique contributions to the understanding and interpretation of video data.

2.7.1 Visual Features

Visual features are fundamental for analyzing and interpreting video data in both general and clinical contexts. These features include color and texture properties [48, 169], which are widely used in traditional video analysis, as well as features extracted from pre-trained deep learning models [170], which capture essential information about the content and context of video frames. In general applications, visual features contribute to tasks such

as object detection [171] and scene segmentation [172], enabling robust and versatile video understanding. In clinical settings, visual features take on an equally critical role, but with a focus on specialized tasks. They are essential for identifying abnormalities, such as lesions or polyps [173], and tracking surgical tools [156]. These applications demand a high degree of precision and context awareness, highlighting the importance of tailoring visual feature extraction to meet the unique requirements of clinical video analysis.

Color and Texture Features

Color features represent the distribution and intensity of colors within a video frame [174]. These features can provide valuable insights into the state of tissues and organs, helping to distinguish between healthy and abnormal conditions. Common techniques for extracting color features include color histograms [174], which visualize the frequency of each color in a given frame, allowing for the identification of specific color patterns that may indicate various tissue types or abnormalities. Another technique is color moments [174], which provide statistical measures of color distribution, such as mean, variance, and skewness, capturing color characteristics in a more compact form. Color features are particularly useful in scenarios such as endoscopic examinations [48], where color changes can signal the presence of lesions or infections.

Texture features describe the pattern properties of objects within the video frames [48], capturing patterns that characterize the structure of tissues. These features are crucial for understanding the texture of surfaces, which can be indicative of pathology. Common methods for extracting texture features include Local Binary Patterns [175], which encode the local texture information by comparing each pixel with its neighbors, providing a robust representation of local textures. Another common method is the Gray Level Co-occurrence Matrix [175], which analyzes the spatial relationship between pixels of different intensity levels, allowing for the extraction of texture descriptors such as contrast, correlation, energy, and homogeneity. Texture features are particularly valuable in identifying conditions like fibrosis or other structural changes in tissues.

In addition to these traditional methods, more advanced local descriptors, such as Scale-Invariant Feature Transform (SIFT) [176] and Speeded-Up Robust Features

(SURF) [177], have been developed to capture distinctive texture patterns. SIFT identifies and describes keypoints in an image that are invariant to scale, rotation, and illumination changes by detecting extrema in Difference-of-Gaussians space and computing gradient-based descriptors. SURF improves upon SIFT by using an approximation based on the Hessian matrix for keypoint detection and Haar wavelet responses for feature description, which significantly accelerates the computation while maintaining robustness to transformations. These methods have been widely adopted in medical imaging applications for tasks such as structure tracking.

Despite their utility, color and texture features face key limitations, particularly regarding variability in recording quality and the fixed nature of their extraction methods. Fluctuations in lighting conditions and image clarity can adversely affect the accuracy of these features [178], leading to inconsistent tissue representation and potential misinterpretation of clinical conditions. Additionally, because the extraction methods for these features are not adaptable or powerful enough for specific tasks [179], they lack the flexibility needed to optimize performance in complex distributions or patterns across diverse clinical scenarios, which often degrades their effectiveness in various analytical frameworks.

Deep Learning Visual Features

Deep learning methods have significantly enhanced the extraction of visual features by enabling models to learn hierarchical and task-specific representations directly from raw video data. Unlike traditional handcrafted features, deep learning models automatically learn multi-level abstractions, as shown by a visualization study [180] that analyzes activations at different layers. These results reveal that early layers typically capture low-level patterns such as edges, corners, and textures, while deeper layers encode more complex structures, including object parts, shapes, and semantic relationships.

These hierarchical features are typically learned through deep neural network architectures such as CNNs (also mentioned in our Section 2.2.1). Popular models, including ResNet [181], VGGNet [182], Inception networks [183], and DenseNet [184], have been widely used for visual feature extraction. These models learn spatial hierarchies of features through stacked convolutional operations and pooling layers. For example, ResNet

introduces residual connections that improve the learning of high-level features in very deep networks, Inception modules allow multi-scale feature extraction by combining different convolutional filters within the same layer, and DenseNet improves feature reuse and gradient flow by densely connecting each layer to every other layer.

More recently, ViTs [5] mentioned in our Section 2.2.3 and their derivatives, such as DeiT [185] and Swin Transformer [186], have offered an alternative way of learning features. Instead of relying on local convolutional operations, these models divide images into patches and apply self-attention to capture global contextual relationships between regions. DeiT is a data-efficient variant of ViT that achieves strong performance without requiring extremely large datasets, while Swin Transformer introduces a hierarchical structure with shifted windows to efficiently model both local and global patterns. As a result, transformer-based models can learn richer long-range dependencies and offer strong global feature representations.

These models are typically pretrained and benchmarked on large-scale datasets such as ImageNet [187], which contains millions of labeled natural images across diverse categories. Benchmarking on such datasets allows for standardized evaluation and comparison across architectures. The learned features serve as robust general-purpose visual representations and are widely transferred to a variety of downstream tasks.

Despite their success, transferring these features to clinical videos poses challenges. Clinical imagery often differs substantially from natural scenes, and the presence of artifacts such as smoke in minimally invasive surgery videos [27] or specular highlights in endoscopic footage [13] can introduce noise into the extracted feature maps. These artifacts may degrade model performance in downstream tasks like action recognition [188]. Therefore, effective deployment in clinical applications often requires fine-tuning on domain-specific datasets or additional adaptations to address clinical data characteristics.

2.7.2 Geometric Features

In computer vision, geometric features refer to quantitative representations that describe the shape, size, and spatial relationships of objects within a given environment [189, 190]. Despite advancements in machine learning and deep learning methods for video analysis, challenges remain in clinical settings, particularly in complex environments

and scenarios involving occlusions [26]. By providing structured information about the spatial relationships between semantics and RGB frames, geometric features can improve various tasks.

These geometric features can generally be categorized into 2D and 3D features [191]. 2D geometric features include bounding boxes captured by object detection [192] and segmentation maps produced by semantic segmentation [193]. 3D features encompass depth maps captured by depth estimation [14] and point clouds generated by 3D reconstruction [194]. Additionally, keypoint detection, including keypoints for anatomical structures [195] and joint points detected through human pose estimation [15], could exist in either 2D or 3D, depending on the design of the landmarks.

2D Geometric Features

2D geometric features are essential in computer vision as they provide critical information regarding the spatial layout and relationships of objects within images or video frames [196]. Although they do not reflect the real-world coordinates of objects, they can capture direct information from the cameras, serving as an important cue for end-users to understand the real 3D world [197].

These features are particularly useful in clinical video analysis, where accurate detection and tracking of instruments, anatomical structures, or abnormalities are vital for effective diagnostics and treatment [198–200]. Moreover, 2D geometric features help users understand complex procedures more effectively [156], assisting learners in clinical training. For instance, precise identification and localization of objects within a surgical video can significantly enhance the efficiency and safety of procedures while also providing valuable feedback for trainees developing their skills [201].

One significant application of 2D geometric features is keypoint detection, which involves identifying specific points of interest in an image that correspond to critical landmarks or features [202]. In clinical applications, keypoints can represent various elements, such as the joints of a human body in pose estimation [66] or specific features on surgical instruments [203]. For human body 2D pose estimation, most clinical applications utilize established pose estimation pipelines, such as OpenPose [204], which employs a multi-stage CNN architecture to detect body keypoints by first generating a coarse

heatmap for each joint and then refining these predictions through successive stages to produce accurate 2D coordinates of human skeletal joints. For other keypoint detection tasks, manual annotation is often necessary to retrain the model for specific applications [203, 205]. A commonly used tool is DeepLabCut [206], which supports both 2D and 3D pose estimation. DeepLabCut utilizes transfer learning to adapt a pre-trained CNN model to specific keypoint detection tasks, allowing for efficient annotation and improved accuracy. However, due to the challenges associated with manual annotation for obtaining 3D coordinate annotations, clinical applications tend to focus primarily on 2D keypoint detection [203, 205]. Despite their utility, the interpretation of keypoints can be subjective for human annotators [207], leading to inconsistencies in labeling when creating new datasets, which may affect the model's performance and reliability in clinical settings.

Bounding boxes are another key component of 2D geometric features. These rectangular annotations specify the location and size of objects within an image or video frame [208]. Bounding boxes are widely used in object detection tasks, enabling the identification and tracking of relevant items in clinical settings, such as surgical instruments [192] or anatomical structures [209]. Typically, object detection models for bounding boxes are trained using existing architectures, provided they are fine-tuned with datasets specific to the required clinical scenarios. One popular architecture for bounding box detection is You Only Look Once (YOLO) [210], which operates by predicting bounding boxes and class probabilities directly from full images in a single evaluation. YOLO's speed and efficiency make it particularly suitable for real-time applications, which is crucial in clinical scenarios. Since annotating objects only requires a few clicks to select a rectangle [211], this method is easy to implement, making it widely used in clinical settings [212]. Additionally, as bounding boxes contain the size of objects in the video frames, they could provide some indication of the distance from the camera [132], which contributes to a more comprehensive understanding of the 3D spatial relationships compared to keypoint detection.

Segmentation maps are also a critical aspect of 2D geometric features. These maps partition an image into multiple segments or regions [213], assigning a label to each pixel based on the object it represents. This technique is essential for the precise delineation of structures in clinical images, allowing for improved visualization and analysis of

pathological conditions or surgical areas [214]. Similar to object detection, segmentation for clinical applications typically requires retraining existing segmentation models on the given clinical dataset [214]. One commonly used model is U-Net [7], which is a symmetric convolutional network that encodes input images into a latent feature space and then decodes these latent features into segmentation masks. However, despite providing more precise location information for objects, the dense pixel-level annotations required for segmentation make it less accessible for deployment [215] compared to object detection. Additionally, segmentation accuracy is often affected by various low imaging quality factors [216], such as frequent occlusions and varying resolutions, which are common in clinical settings.

While these 2D geometric features are important for clinical video analysis in describing abstract relationships in the video, many clinical procedures involve complex actions in a 3D coordinate system, such as interactions between clinical trainees and training environment [217]. Thus, while 2D features efficiently represent abstract relationships and structured understanding in the recorded frames, they may sometimes be insufficient if the task requires more understanding in the real 3D coordinate system.

3D Geometric Feature

3D geometric features are crucial for enhancing the understanding of spatial relationships in video analysis, particularly in clinical settings. Unlike their 2D counterparts, which provide limited information about the depth and the volumetric structure of objects, 3D features offer a more detailed representation, enabling precise spatial mapping of movements and relevant objects within the video [218]. By capturing three-dimensional aspects such as depth [14] and spatial location [15], these features facilitate a more accurate interpretation of clinical scenarios.

Keypoints could be captured in a 3D format to provide spatial awareness that closely resembles real-world coordinates. This is commonly used for human skeleton data pose estimation in clinical applications. Existing solutions for 3D joint keypoint estimation include algorithms integrated into depth cameras, such as the Azure Kinect, which provide 2D or 3D coordinates of human skeletal joints [219]. Moreover, pipelines such as MediaPipe Hands [220] could infer 3D hand joints at high speed from single-view live

video. The efficient representation of spatial movement for clinicians or patients makes the use of 3D keypoints in human skeleton data very common in rehabilitation [155].

Additionally, depth maps are a crucial form of 3D information that represent the distance from the camera to various objects in a scene [221]. This information is essential in clinical video analysis, as it enables the reconstruction of three-dimensional structures from two-dimensional images, enhancing spatial understanding during procedures such as endoscopy or surgery. Depth maps are commonly utilized in clinical settings because they can either be directly estimated from raw monocular videos or captured using external devices, making them highly suitable for clinical applications. Due to the lack of ground truth data, self-supervised learning has emerged as a prevalent method for monocular depth estimation in clinical contexts, addressing issues of accessibility [137]. This approach leverages disparity information [14], which refers to differences in pixel positions between two views of the same scene. By utilizing warping-based view synthesis and simultaneous depth and ego-motion estimation, the appearance differences between the target and synthesized frames serve as a supervisory signal for training. Thus, depth maps can be considered an accessible solution for depth estimation in clinical settings, as they can be generated without requiring extensive annotations..

Another important 3D geometric feature is the point cloud, which consists of a collection of data points in three-dimensional space, each representing the surface of an object or scene [222]. Point clouds could be generated through techniques such as 3D reconstruction using 3D scanners or LiDAR [222], enabling detailed spatial mapping of clinical environments. However, these methods typically require additional hardware, which limits their application in standard inner-body scenarios such as surgery, where additional cameras or sensors are generally not installed [108]. These limitations reduce the feasibility of directly using point clouds in clinical settings.

In addition to direct point cloud capture, alternative approaches attempt to construct dense 3D representations from multi-view frames without the need for external devices, often leveraging Structure-from-Motion (SfM) techniques [223]. SfM uses keypoint matching across frames to estimate camera poses and reconstruct the 3D geometry of a scene. While these methods are hardware-independent, they rely heavily on accurate camera pose estimation and reliable keypoint tracking to effectively combine data across

frames. Unfortunately, in clinical settings, challenges such as occlusion, limited fields of view, and difficulties in maintaining stable keypoints on textureless surfaces often lead to inaccuracies in pose estimation [224]. These issues significantly limit the application of SfM in clinical settings, particularly in surgical and endoscopic scenarios.

Similarly, volumetric scene representations have recently gained popularity as a 3D geometric feature. Neural Radiance Fields (NeRF) has emerged as a novel method for synthesizing complex 3D scenes from 2D images by modeling volumetric scene representations [225]. Recent efforts have aimed to adapt this technology for clinical applications [226, 227]. For example, Gerats *et al.* [227] use NeRF to reconstruct 3D scenes of surgical procedures from sparse-view RGB-D videos, combining time-of-flight sensor data and dense depth estimations to create geometrically accurate visual representations with fewer cameras. However, NeRF also requires accurate camera pose estimation and typically demands significant computational resources and time for inference. It is not designed for real-time applications [228], which limits its widespread use in clinical video analysis.

Recently, 3D Gaussian Splatting (3DGS) [229] has emerged as a promising alternative to NeRF for 3D scene reconstruction. Instead of modeling a continuous volumetric field through neural networks, 3DGS represents a scene using a set of anisotropic Gaussian functions directly placed in 3D space, allowing for efficient real-time rendering without expensive ray marching. Compared to NeRF, 3DGS achieves faster training times and significantly higher rendering speed while maintaining competitive visual quality. However, despite its advantages, 3DGS also has limitations. It typically requires accurate camera poses and high-quality multi-view images for effective reconstruction, which are difficult to obtain in clinical settings where camera movement is restricted and views are often sparse or occluded. Moreover, the representation assumes static scenes, making it less suitable for highly dynamic environments such as surgeries or endoscopic examinations [10]. As a result, while 3DGS offers SOTA performance in general 3D reconstruction, its application to clinical video analysis remains challenging and largely unexplored.

2.7.3 Justification of Feature Selection in This Thesis

In this review, both visual and geometric features are explored for clinical video analysis. While visual features, including color, texture, and deep learning-based representations, provide semantic information about the appearance of clinical scenes, their effectiveness often diminishes in the presence of noise, lighting variation, and occlusion [13]. To address these challenges, this thesis emphasizes geometric features, which encode structured spatial information about the scene and complement visual features with enhanced robustness and interpretability.

However, not all geometric features are equally suitable for clinical applications. Complex 3D representations such as volumetric reconstructions [225, 229] or dense point clouds often require multi-view hardware setups or high-quality camera pose estimation. These conditions are difficult to meet in real-world clinical environments. On the other hand, simpler 2D representations like keypoints may lack sufficient spatial context. Thus, the selected features are carefully chosen based on their task relevance, practicality, and balance between 2D and 3D information.

Bounding boxes are selected for the surgical workflow anticipation task (Chapter 3) due to their strong performance in object localization and tracking while maintaining simplicity and ease of annotation. Compared to keypoints or segmentation masks, bounding boxes provide a structured and interpretable way to model temporal object interactions without requiring detailed annotations [208]. For long-term anticipation tasks, this lightweight representation enables efficient modeling of tool trajectories and object presence over time, which are critical cues for predicting future procedural steps.

Depth maps are used for endoscopic video inpainting (Chapter 4) because they offer accessible 3D structural information that enhances visual recovery in degraded frames. Unlike surface-level 2D features or expensive volumetric methods, depth maps could be estimated directly from monocular video using self-supervised learning [14], avoiding the need for special hardware. This makes them highly practical for clinical enhancement, where occlusions, lighting artifacts, and fluid interference often obscure essential anatomical details. Integrating depth helps guide the reconstruction toward spatially consistent and clinically plausible content.

3D skeleton data is employed for clinical skill assessment (Chapter 5) because it

captures detailed motion patterns that are critical for evaluating procedural accuracy. Compared to bounding boxes, skeletons provide explicit joint-level representations of the hands and body [219], which are necessary for analyzing fine motor skills such as needle manipulation in acupuncture or chest compression techniques in CPR training. This level of detail enables more precise tracking of skill execution. Furthermore, 3D skeletons support multi-view motion modeling, improving robustness to occlusion and allowing consistent evaluation from different camera angles, which is essential in clinical training environments.

In summary, this thesis selects bounding boxes for temporal workflow modeling, depth maps for spatial restoration, and 3D skeletons for motion-level analysis. These choices reflect a task-specific strategy to balance precision, computational efficiency, and practicality in diverse clinical video scenarios.

2.8 Geometric feature enhanced Deep Learning for Video Analysis

Geometric features significantly enhance the effectiveness of deep learning models in clinical video analysis by providing a more structured understanding of spatial relationships and interactions within the videos [21,54]. In contrast to purely visual features, geometric features, such as bounding boxes, depth maps, and key points, offer valuable spatial information that enables a deeper interpretation of clinical events, instrument interactions, and the performance of clinical tasks. This section explores three key areas where geometric feature enhanced deep learning is making a substantial impact: long-term anticipation, video quality enhancement, and fine-grained semantic understanding.

2.8.1 Geometric Feature Enhanced Long-term Anticipation

In long-term video anticipation, geometric features such as bounding boxes are valuable for capturing spatial relationships that purely visual features may overlook. A common approach involves utilizing existing object detection models to generate bounding boxes for objects and embedding these geometric features into neural network layers to enhance

the overall framework [21, 132, 230]. In general computer vision, for example, Furnari *et al.* [230] use bounding boxes to represent object occurrences, which are then fused with visual features through an attention mechanism to anticipate the next action.

In clinical settings, bounding boxes have also been introduced to enhance anticipation frameworks. For example, Yuan *et al.* [21] propose Instrument Interaction Aware Anticipation Network (IIA-Net), which improves surgical workflow anticipation by incorporating geometric features such as bounding boxes to represent interactions between instruments. These features are encoded and fused with visual information to enhance the model's ability to predict surgical steps.

Despite these advancements, their method focuses primarily on instrument bounding boxes, limiting its ability to track the movement of anatomical structures and neglecting critical interactions between instruments and anatomy. Incorporating more detailed geometric features, such as bounding boxes of anatomical structures, could provide a more comprehensive understanding of the surgical context and significantly enhance the accuracy of event predictions.

2.8.2 Geometric Feature Enhanced Video Quality Improvement

Geometric features, such as depth information, are critical for improving video quality tasks. In general video inpainting, depth maps are employed to reconstruct missing or occluded parts of a video by embedding depth features, concatenating them with RGB visual features, and using the combined depth-enhanced features to guide the inpainting process [231, 232]. However, many methods rely on specialized equipment, such as LiDAR [231], or require precise depth ground truth during inference [232], which is often impractical in clinical settings, particularly for endoscopic video capture [108].

In clinical scenarios, some attempts have been made to leverage depth information. For instance, Chen et al. [52] proposed a framework for improving endoscopic video resolution using depth information, focusing solely on texture refinement relative to depth ranges and single-image processing without incorporating temporal information. This limitation hinders the framework's ability to effectively address spatial structure understanding in videos, reducing its applicability for improving endoscopic video quality.

To the best of our knowledge, no other methods besides the approach introduced in

Chapter 4 have utilized depth information specifically for video inpainting in endoscopic settings. Given the critical need for real-time video analysis in endoscopy, existing geometric-based methods relying on depth maps often require customized solutions to effectively address the unique challenges posed by these environments.

2.8.3 Geometric Feature Enhanced Fine-Grained Semantic Understanding

Geometric features, such as human skeleton data, which represent key points corresponding to joints and body movements, have become essential inputs for skill assessment frameworks requiring fine-grained semantic understanding [150]. Early clinical training systems directly modeled skeleton sequences using RNNs. For example, STNN [22] employs a LSTM [97] network to assess physical rehabilitation exercises based on human pose data.

To better capture both spatial and temporal dependencies in human motion, Graph Convolutional Network (GCN)-based methods have been proposed. Unlike traditional convolutional networks that operate on regular grids such as images, GCNs generalize convolution operations to graph-structured data [233], where nodes represent entities (e.g., joints) and edges represent relationships (e.g., bones or connections between joints). By aggregating information from neighboring nodes, GCNs are well-suited to model the spatial structure of the human body. Building upon this idea, Spatial-Temporal Graph Convolutional Network (ST-GCN) [234] extend GCNs by introducing temporal edges across consecutive frames, allowing the model to jointly learn spatial relationships within a frame and temporal dynamics across frames. This enables structured and efficient modeling of joint movement patterns over time. Many recent clinical training systems adopt ST-GCN-based architectures for fine-grained semantic understanding in clinical applications [158, 235]. For example, STGCN-LSTM [158] integrates an LSTM module into the ST-GCN framework to model long-term temporal patterns for rehabilitation assessment, while STGCN-RI [235] further enhances the ST-GCN by incorporating joint rotation matrices to improve robustness in skeleton-based skill evaluation.

Despite these advantages, these frameworks often rely on single-view data [158, 235], which limits their ability to address occlusions commonly encountered in clinical practice,

thereby hindering comprehensive movement analysis. Furthermore, as these methods primarily focus on the skeleton of the practitioner, they overlook critical interactions with the environment, such as equipment or other trainees [158, 235]. Therefore, by relying solely on single-view skeleton data, these frameworks may fail to provide sufficient context for evaluating complex movement patterns and assessing how effectively the trainee interacts with their surroundings, which are essential components of clinical performance.

In summary, geometric feature-enhanced deep learning holds significant potential for advancing clinical video analysis. From long-term video modeling to video enhancement and fine-grained semantic understanding, the integration of geometric features provides a richer and more structured interpretation of clinical scenarios. However, there remains considerable scope for improvement, particularly in capturing the interactions between instruments and anatomy, developing specialized solutions for enhancing endoscopic video quality with depth assistance, and achieving a comprehensive multi-view understanding of skill interactions within clinical environments. To address these gaps, this thesis proposes multiple geometric feature enhanced clinical video analysis frameworks. Specifically, it introduces structured object-location modeling for improving long-term surgical workflow anticipation (Chapter 3), depth-guided video inpainting for robust endoscopic video quality improvement (Chapter 4), and multi-view human-object interaction modeling using skeleton features for fine-grained clinical skill assessment (Chapter 5).

CHAPTER 3

Surgical Workflow Anticipation Leveraging Geometric Features

Portions of this chapter have previously been published in the following peer-reviewed publications:

- **Zhang, F. X.**, Moubayed, N. A., & Shum, H. P. H. (2022). Towards graph representation learning based surgical workflow anticipation. In *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI '22)*, pages 1-4, IEEE.
- **Zhang, F. X.**, Deng, J., Lieck, R., & Shum, H. P. H. (2025). Adaptive graph learning from spatial information for surgical workflow anticipation. *IEEE Transactions on Medical Robotics and Bionics*, *7*(1), 266–280.

In this chapter, we explore how geometric features could enhance long-term video anticipation in clinical settings, focusing on the challenging task of surgical workflow anticipation. This task involves managing extended time scales and interpreting abstract interactions between surgical instruments and anatomical structures [21], making it a

compelling validation case for long-term video understanding in clinical applications. As the opening chapter of this thesis, and in contrast to subsequent sections (Chapter 4 and Chapter 5) that explore the fusion of geometric and visual features, this work emphasizes the potential of geometric features as standalone robust representations for complex clinical scenarios, without relying on fusion with other features. It demonstrates their ability to support deep learning models in capturing long-term dynamics in clinical videos.

To address this challenge, we propose a surgical workflow anticipation method that uses geometric features of both surgical instruments and anatomical targets as primary inputs. Extending earlier work focused solely on instruments, this chapter incorporates both instruments and targets to model complex interactions. An adaptive graph method dynamically selects optimal graph representations for each video frame, while a multi-horizon objective balances learning across time horizons, enabling flexible predictions. Evaluations on two benchmarks show a 3% error reduction in surgical phase anticipation and 9% in predicting surgical duration, demonstrating improved short-to-mid-term anticipation. These results highlight the method's potential to enhance surgery team coordination and emphasize the role of geometric features in advancing long-term clinical video analysis.

3.1 Introduction

Surgical workflow anticipation is the task of automatically predicting the timing of relevant surgical events from live video data, such as the remaining time before a surgical instrument is changed or the remaining duration of the entire surgery. The capability of this task facilitates efficient instrument preparation and the design of intelligent robotic assistance systems [4]. Additionally, it can enhance patient safety and facilitate communication in the operating room [74]. Consequently, using machine learning models to improve surgical workflow anticipation has become an important research topic in surgical vision [4, 133, 134, 136] and is critical for the efficiency and safety of Minimally Invasive Surgery (MIS) [21].

To make accurate surgical workflow predictions, it is crucial to consider the interac-

tions between surgical instruments and targets (e.g., a gripper fixating tissue) [73], as these directly influence subsequent steps and surgical outcomes. This process requires integrating geometric features [236], with a particular emphasis on tracking changes in location and size over time. Such tracking helps describe the rational movement and interaction relationships between surgical instruments and targets [237]. Recent approaches [21] utilize these geometric features, which are easily captured by existing object detection models, and demonstrate robustness even during frequent partial occlusions in complex surgical environments [199] (as shown in Fig. 3.1, where object detection could still detect robust geometric features for surgical instruments and targets during complex interactions). Similarly, in our initial attempt for this task [135], we also used the bounding box of surgical instruments as our primary input and designed a graph convolution to model the relationships between these instruments, showing a significant improvement in short-term anticipation. However, previous works and our early attempts often did not give enough attention to tracking the surgical target, which misses the important interactions between instruments and surgical targets. This limitation restricts the potential for a comprehensive representation of surgical scenarios.

Anticipation models are generally trained and evaluated based on their prediction accuracy as measured by the Mean Absolute Error (MAE). In practice, immediately impending events are more important than those in the remote future. In the surgical workflow anticipation field, it is therefore common to define a so-called *time horizon* h (*e.g.,* 2min/3min/5min) [4, 21], which sets a temporal threshold to differentiate between *in-horizon* events ($t \leq h$) and *out-of-horizon* events ($t > h$). These two types of events are evaluated differently [4]: in-horizon events occur within the given time horizon and should be predicted as accurately as possible, whereas out-of-horizon events occur beyond the given time horizon and should be largely ignored by setting the error to zero as long as the model predicts any time greater than h. The reason for setting the error to zero is to avoid penalizing the model for imprecise predictions of events that are too far in the future to be clinically actionable [238]. This design focuses the evaluation on near-future anticipation accuracy, which is more relevant for real-time decision-making in surgical settings. However, previous works [4, 21] typically use a fixed duration for h (*i.e.,* a *fixed horizon* [239]). This approach limits the model's ability to

adapt to varying surgical durations because it cannot optimize for different time horizons within a single framework (*i.e., multi-horizon* modeling [135]). As the length of surgery can vary depending on the patient's condition [133], relying on a fixed time horizon restricts the model's generalizability across a broader patient population.

We identified three major challenges for employing geometric features for surgical workflow anticipation. First, existing methods provide an incomplete and unreliable representation of the surgical scene [21, 135]. Their representation is limited to surgical instruments, which ignores surgical targets as well as the uncertainty associated with the process of geometric feature extraction. This lack of comprehensive information limits their ability to represent interactions effectively. Second, previous methods [205] and our early attempts [135] use static graphs that cannot capture the dynamic interactions between instruments and surgical targets, which may vary significantly across the surgery. For example, in the cholecystectomy shown in Fig. 3.1, the graspers are initially used to position the gallbladder for broad exposure. As the procedure progresses, the hook joins in to separate the gallbladder from the liver, necessitating a different interaction representation for the current frames. These changes in interaction cannot be captured by static graphs and require adaptive graphs. Third, existing anticipation methods, including those based on geometric approaches [21, 135], struggle with the diverse time span requirements of surgical anticipation. They typically employ a fixed time horizon in their model training and evaluation. When training, any prediction times exceeding the given time horizon are adjusted back to the fixed time horizon value. This approach limits their applicability to scenarios with a fixed horizon, while the time scales of surgical scenarios can vary significantly [133].

To address these challenges, we present an adaptive graph learning framework for surgical workflow anticipation based on a novel geometric representation. First, to represent the surgical scene more comprehensively, we propose a new geometric representation based on bounding boxes. This representation includes both surgical instruments and targets, along with their detection confidence levels. Due to the lack of annotations for surgical targets in popular benchmark datasets, namely the Cholec80 cholecystectomy video dataset [2] and the Cataract101 cataract surgery video dataset [162], we provide additional annotations for surgical targets. These annotations enable us to train object

detection models for both surgical instruments and targets. Second, we introduce an adaptive graph learning approach to promote a dynamic understanding of the surgical procedure. Unlike our prior static methods for graph selection [135], our method dynamically selects suitable candidate graphs for each frame in the video thus adapting the graph representation over time according to the current surgical situation. Third, to meet the diverse time horizon requirements of complex surgical settings, we introduce a multi-horizon objective that combines the loss functions for different time horizons using learnable weights. This optimizes a generalized prediction across different horizons without manual weight adjustment.

Comprehensive experiments on two benchmark datasets indicate that our method outperforms existing surgical anticipation methods. To promote a fair comparison with previous methods, we employed the commonly used variants of MAE as an evaluation metric, where a higher MAE implies inaccurate predictions that may delay the timely handling of relevant events and adversely affect patient outcomes [4]. Specifically, compared to SOTA method proposed by Yuan et al. [21], our approach achieves a $\sim$3% reduction in MAE for predicting the beginning of the next surgical phase (*surgical phase anticipation*) and a $\sim$9% reduction for predicting the end of the surgery (*remaining surgery duration anticipation*) compared to existing benchmarks. This highlights the potential of our method for enhancing preparation and coordination within the surgery team, which helps improve surgical safety and the efficiency of operating room usage.

A demonstration of our framework running on live surgical video can be found in the shared video[1]. Our main contributions are as follows:

1. We propose a novel geometric representation based on bounding boxes to extract geometric features about both instruments and surgical targets, along with their detection confidence levels. To train our object detection models and address the current gap in surgical target detection, we provide additional object annotations in two benchmark datasets.

2. We introduce an adaptive graph learning approach to dynamically select a number of candidate graphs for the current time frame that represent interactions

[1] https://www.youtube.com/watch?v=kme1_oJsyeA

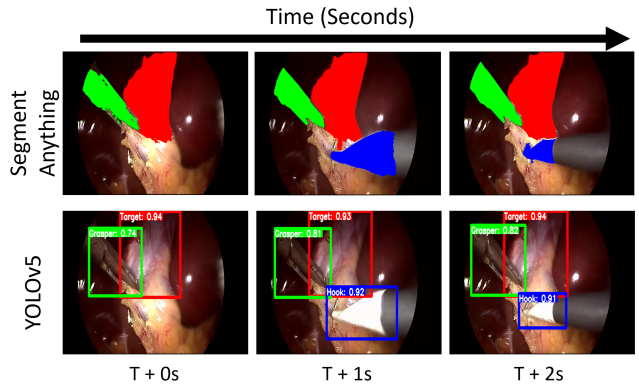

Figure 3.1: Comparison of semantic segmentation (**Top**) with object detection (**Bottom**) across three consecutive seconds: We compare our trained YOLOv5 model with Segment Anything [11], a state-of-the-art foundation model for segmentation. Segmentation masks significantly change across frames even when their positions remain static. In contrast, bounding boxes consistently provide a stable representation of both location and size.

among surgical instruments and targets. With graph convolution and temporal convolution, this allows for leveraging dynamic geometric features to improve our predictions in complex surgical settings.

3. We design a multi-horizon training strategy that incorporates loss terms for multiple time horizons, thus eliminating the need for choosing a single, fixed horizon. Our multi-horizon objective allows for automatically balancing the terms for different horizons, thereby making our approach more flexible and broadly applicable.

3.2 Method Overview

Our method uses a series of raw video frames as input and predicts the timing of various surgical events as output. It consists of three main processing stages, depicted in Fig. 3.2 and described in more detail in Sections 3.3–3.5.

Geometric Feature (Section 3.3)**:** In the first stage, we extract bounding boxes from the raw video frames using YOLOv5 [240]. By providing essential geometric features (*i.e.,* location and size) with less complexity, bounding boxes offer a more stable geometric representation compared to pixel-level segmentation. At the same time, they provide all the necessary geometric features required for subsequent steps. To further improve the

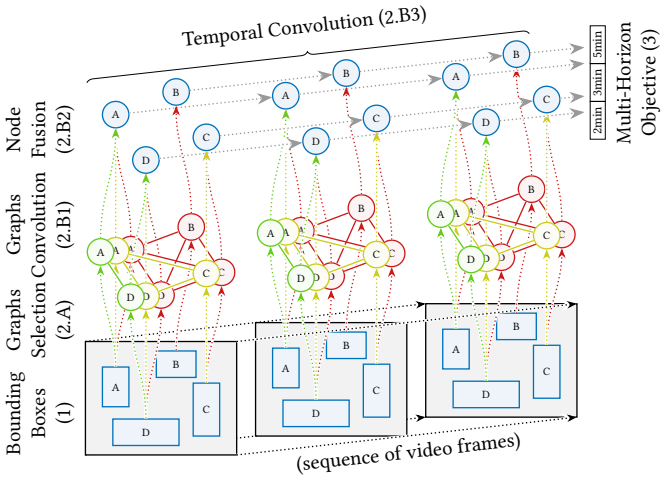

Figure 3.2: Overview of our method. Given a sequence of video frames as input, our model has three main processing stages (Sections 3.3–3.5): (1): From the raw frames, we extract bounding boxes of surgical instruments and targets. (2): This information is further processed using adaptive graphs by (2.A) selecting a number of candidate graphs (red, yellow, green), (2.B1) using graph convolution to process the node features based on the graph's connectivity; (2.B2) fusing nodes from the multiple candidate graphs, and (2.B3) performing temporal convolution over the nodes from different video frames. (3): The final node features are used to produce an unconstrained prediction of various surgical events trained using a multi-horizon objective.

quality and robustness of the basic YOLOv5 model, we created additional annotations (Fig. 3.3) and fine-tuned YOLOv5 on these datasets (3.3.1). Moreover, we employ confidence estimates as a quantifiable reference of potential uncertainty in the geometric feature extraction (3.3.2).

Adaptive Graphs (Section 3.4)**:** A core contribution of our work is the use of adaptive graphs to model the interaction between surgical instruments and targets over time. This involves two phases, first, the selection of candidate graphs based on both prior statistics about typical interactions as well as geometric features that are dynamically extracted from the bounding boxes (3.4.1), and, second, further processing of the node features through graph convolutions, node fusion, and temporal convolutions (3.4.2).

Multi-Horizon Objective (Section 3.5)**:** It is common to train models for only predicting events within a fixed time horizon h (2min/3min/5min) using dedicated loss functions. In contrast, our model is trained to predict the actual timing of events, independently of any artificially imposed time horizon constraints. When evaluating the model on existing fixed-horizon benchmarks, the model output is clipped to the respective horizon h. For

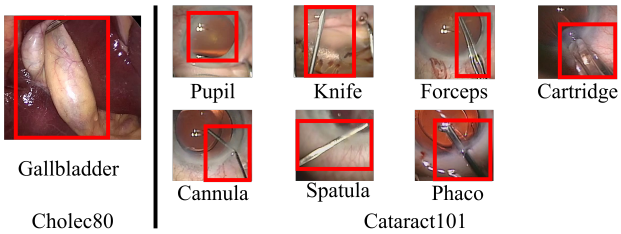

Figure 3.3: Additional annotation for existing datasets. We provide additional annotations for the Cholec80 dataset focusing on surgical targets and for the Cataract101 dataset covering both surgical targets and instruments.

training, we combine the loss functions for multiple horizons with a learnable weighting to a single multi-horizon loss. This allows our model to make unbounded predictions while still putting higher weight on the more relevant short-term predictions.

3.3 Geometric Feature Representation

A major challenge in leveraging geometric features for surgical workflow anticipation [21, 135] is the lack of a stable representation of both instruments and surgical targets. This absence limits their effectiveness in representing critical interactions during surgery. Although previous efforts have utilized segmentations to represent surgical targets [21], this method primarily captures the shape of the surgical scene. Compared to object detection, segmentation often fails to directly describe the motion of instruments and surgical targets, which is crucial for understanding surgical workflows. Additionally, in the highly non-rigid surgical environment [23], the shape of estimated segmentation masks frequently changes during interactions between instruments and targets. This variability often introduces noise into the training process, as demonstrated in the top part of Fig. 3.1.

To address this challenge, our framework employs bounding boxes to represent both instruments and surgical targets. It provides a stable geometric representation by focusing on key characteristics such as location and size, which remain consistent during interactions between instruments and targets, as illustrated in the bottom part of Fig. 3.1.

3.3.1 Additional Data Annotation

There are two significant issues with existing datasets. First, they do not include annotations for surgical targets [162, 192]. Second, most annotations are on selected high-quality images. This results in models being susceptible to image artifacts, such as reflection and motion blur, that can impair object detection performance.

To address these issues, we provide additional bounding box annotations on existing datasets for *both* instruments *and* surgical targets. Importantly, our annotations include a variety of image qualities without specifically selecting for high quality. This allows object detection models trained on our dataset to better handle the artifacts presented in real-world surgical settings. The annotations were performed by the PhD candidate (the author of this thesis), who holds a medical bachelor's degree, providing relevant domain knowledge to guide the annotation process.

Specifically, our additional annotations are based on two popular surgical video datasets: the Cholec80 cholecystectomy dataset [2] and the Cataract101 cataract surgery dataset [162]. Figure 3.3 showcases examples of these additional annotations.

For surgical target annotations, we define surgical targets as the tissue regions most frequently interacted with by instruments, identified based on our observation of the entire video for each surgical case. For the Cholec80 dataset, we provide additional annotations for 5,137 frames, focusing on surgical targets (*i.e.,* the gallbladder and adjacent tissues targeted for removal). For the Cataract101 dataset, we provide additional annotations for 2,157 frames of surgical targets (*i.e.,* the pupil).

For instrument annotations, we leverage existing datasets. For Cholec80, we use the m2cai16-tool-locations dataset [192]. For the Cataract101 dataset, we provide additional annotations for 2,990 frames, employing a simplified instrument definition adapted from Fox et al. [241].

We employ YOLOv5 [240] to extract geometric feature (Section 3.3.2). YOLOv5 is known for its effectiveness in challenging conditions. This is due to its advanced data augmentation techniques that simulate visual distortions and its multiscale architecture, which is capable of capturing objects of various sizes. We additionally fine-tune YOLOv5 on our new datasets to further enhance its object detection robustness. This enables us to accurately detect instruments or surgical targets despite environmental variability,

facilitating effective geometric feature extraction.

3.3.2 Confidence Estimates

Previous anticipation methods [21,135] that leverage geometric features (whether through bounding boxes or semantic segmentation) typically do not account for the uncertainty of their geometric feature. This ignorance often leads to reliance on potentially inaccurate geometric features. Such overconfidence often leads to incorrect information being used for anticipation, particularly in complex surgical scenarios where the extraction of geometric features is noisy.

To address this issue, our approach includes bounding boxes and their detection confidence levels. This integration provides subsequent models with a quantifiable reference of potential uncertainty in the current geometric feature extraction [210]. This method offers improved reliability over models that exclusively use segmentation or bounding boxes.

Specifically, we extract the bounding box features as:

$$b_t = \{x_t, y_t, w_t, h_t, c_t\}\,, \tag{3.1}$$

where x_t and y_t denote the x and y coordinates of the center of a surgical element in frame t, w_t and h_t represent the detected width and height of each surgical element in frame t, and c_t denotes the detection confidence score directly output from the object detection model, indicating the reliability of the bounding box. Including c_t in the geometric representation allows the model to account for the trustworthiness of each spatial input and reduce the impact of noise from potentially unstable detections. The observed sequence $\{b_0 \cdots b_T\}$, spanning from time point 0 to the current observed time point T, is utilized to anticipate when surgical event e occurs. Denoting the number of instruments and surgical targets in the surgery type to anticipate as N and the feature number in the bounding box as B, the observed geometric feature $b_T \in \mathbb{R}^{T\times N\times B}$. To further improve the robustness of our representation against potential inaccuracies, we adopt a learnable weighting for our representation to discern potential reliable frames through a temporal attention mechanism [242]. This mechanism assigns weights to

frames based on object detection results and their associated confidence levels.

3.4 Adaptive Graph Learning

A major challenge for our early attempt [135] employing geometric features is their reliance on a single static graph to describe the interaction relationships between instruments and surgical targets. This approach assumes fixed surgical interactions throughout the surgical procedure and fails to accurately capture the dynamic interactions between various instruments and surgical targets during surgery. Consequently, a single static graph often cannot offer an appropriate representation of interactions for different frames within a surgical video.

To address this challenge for dynamic surgical interactions, we design our anticipation feature learning to leverage adaptive graphs. These adaptive graphs automatically select different suitable representations for each frame, accurately reflecting the dynamic surgical interactions. Our method comprises two main steps: Candidate Graph Selection (Section 3.4.1) which determines the graph representation policy for each frame, and Graph-based Feature Learning (Section 3.4.2) which transforms geometric feature into spatio-temporal features for anticipation based on the selected graphs for each frame, and then transforms the feature representation into the final anticipation output. The procedure is illustrated in Fig. 3.4.

3.4.1 Candidate Graph Selection

Two key issues arise when employing graph selections for surgical interaction analysis. First, the selected graphs should accurately represent surgical interactions [243]. Secondly, it is critical to design a graph selection process that is learnable to optimize it for the specific prediction task.

To address these issues, our candidate graph selection process involves two steps. First, before training, we generate an initial set of fully connected candidate graphs. This set includes the most frequently occurring combinations of instruments and surgical targets, according to object detection results from each frame in the training set. This approach identifies the most representative interactions within the dataset. Second, during

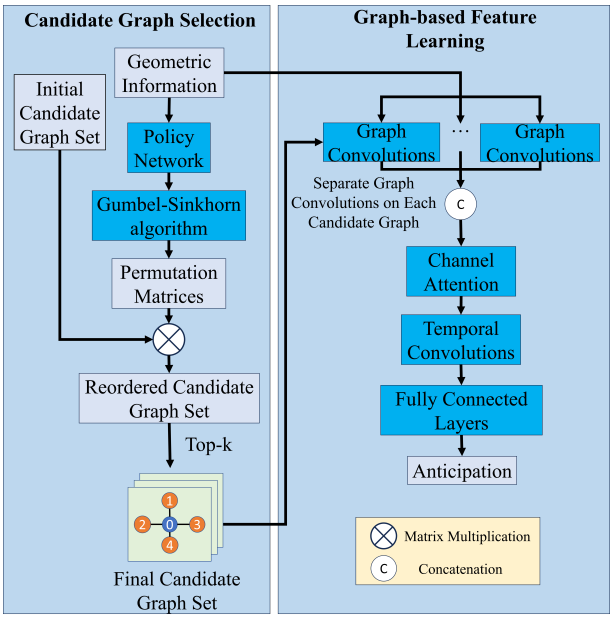

Figure 3.4: The architecture of our adaptive graph learning consists of two main components: **Left**: Candidate Graph Selection selects the suitable graph representations for each frame from the most common interactions observed in the training data. **Right**: Graph-based Feature Learning transforms geometric features into spatio-temporal features for anticipation based on the selected graphs for each frame, and then transforms the feature representation into the final anticipation output.

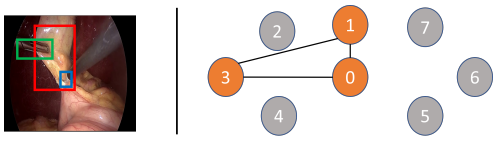

Figure 3.5: Example of object detection results and graph representation from a cholecystectomy [2]. **Left:** A frame showing the grasper and hook dissecting the tissue plane. **Right:** Fully connected candidate graph representing interactions among instruments and surgical targets. Gray nodes represent objects that do not appear in the frame. **Node legend:** 0: surgical target; 1: grasper; 2: bipolar; 3: hook; 4: scissors; 5: clipper; 6: irrigator; 7: specimen bag.

training and inference, we employ a policy network enhanced by the Gumbel-Sinkhorn operation [244] (further explained in Eq. 3.3) to select the most relevant candidate graphs for each frame. The whole graph selection process is end-to-end differentiable, allowing us to optimize it for our anticipation tasks.

Specifically, in our first step, we construct a raw candidate graph set, denoted as G. This set comprises various graphs that encapsulate all possible combinations of instruments and surgical targets, which are identified from object detection results in each frame of the training data. Within these graphs, we define a node set V with N node where N denotes the number of instruments or surgical targets in the current frame. Each node v_i represents a bounding box of a surgical instrument or the surgical target and N denotes the number of instruments or surgical targets in the current frame. The edge set E encapsulates the interactions between these nodes, defined as $E = \{v_i v_j | i, j \in H, i \neq j\}$, where H represents the set of all observed interactions based on the all possible combinations of instruments and surgical targets. Each graph in G is associated with an $N \times N$ adjacency matrix A, where an entry $A^{ij} = 1$ signifies an interaction (*i.e.,* an edge) between nodes v_i and v_j. Figure 3.5 provides a visual example of object detection results and their corresponding graph representation.

To concentrate on the most representative interactions, we select C graphs with the most common combinations according to their frequency of occurrence in the dataset. These form the initial candidate graph set $G_C \in \mathbb{R}^C$. This set is then applied to the sequence of frames in our observed surgical videos, creating a corresponding sequence of initial candidate graph sets denoted as $G_T \in \mathbb{R}^{T \times C}$.

Then, in our second step, we derive a preliminary permutation likelihood matrix sequence $M_T \in \mathbb{R}^{T \times C \times C}$ from $\hat{b}_T$ via a policy network $Policy$, which is a l_p-layer causal TCN [21] with l_p layers. This sequence guides the reordering of G_T:

$$M_T = Policy(\hat{b}_T) \tag{3.2}$$

To convert the likelihoods into a binary permutation matrix but still enabling gradient backpropagation for end-to-end learning, we utilize the Gumbel-Sinkhorn algorithm [244], allowing for differentiable binarization of matrix $M_t \in \mathbb{R}^{C \times C}$ of each matrix from M_T.

This approach first introduces a Gumbel noise term GE_t to inject randomness into the permutation process, promoting that the model explores diverse permutations during training. Then, the Sinkhorn algorithm iteratively refines the matrix towards a doubly stochastic matrix that approximates binary permutation values:

$$\begin{aligned} M_t^0 &= exp(M_t + GE_t) \\ M_t^n &= \tau_c(\tau_r(M_t^{l-1})) \\ \hat{M}_t &= \textit{Softmax}(M_t^n/\tau)\,, \end{aligned} \tag{3.3}$$

where $GE_t = -\log(-\log(U_t))$ and U_t is sampled from a uniform $i.i.d$ distribution. M_t^n denotes the permutation matrix at the n-th Sinkhorn iteration. This iterative process refines M_t^0 to approximate binary values by sequentially applying column-wise normalization τ_c and row-wise normalization τ_r to P_t^{l-1}. $\hat{M}_t \in \{0,1\}^{C\times C}$ denotes the optimized permutation matrix of each frame and forms the permutation matrices $\hat{M}_T \in \{0,1\}^{T\times C\times C}$ for observed frames. τ denotes the temperature parameter adjusting the sharpness of approximation. Balancing computation cost and performance, we set the Sinkhorn iterations to 10.

Finally, we reorder G_T by multiplying it with the optimized permutation matrices $\hat{M}_T$. To focus on the most relevant candidate graphs for each frame, we leverage the top k rows from the reordered matrices. This forms the final candidate graph set for each frame and a sequence of final candidate graph sets $\hat{G}_T \in \mathbb{R}^{T\times k}$ for the observed video frames.

3.4.2 Graph-based Feature Learning

A key limitation of existing graph-based feature learning in surgical workflow anticipation is the application of a fixed single graph and its graph convolutions to every frame in a surgery video [135]. This approach cannot accommodate the need for diverse graph convolutions to update graph node features based on varying surgical interactions because of the different graph representations selected across frames.

To overcome this limitation, we design a specific graph-based feature learning process that effectively incorporates selected candidate graphs. First, it employs independent

graph convolutions for each selected candidate graph. This method promotes tailored updates to node features according to the distinct characteristics of different graphs. Second, we introduce a channel attention mechanism [242] to efficiently fuse node features across various graphs, providing the flexibility to integrate diverse graph combinations. Third, we incorporate dilated causal 2D convolutions to effectively aggregate current node features with those from previous frames. This embeds a broad temporal dimension into the feature learning to capture long-term temporal correlations.

In particular, first, each graph $\hat{G}_{T_k}$ from $\hat{G}_T$ is processed through 2-layers of geometric graph convolutions to encode geometric relationships and inject more semantic information into input geometric feature [233] where $\tilde{H}_0 = \hat{b}_T$:

$$\tilde{H}_{l_g}^{(s)} = \Lambda_k^{-1/2}(A_k + I)\Lambda_k^{-1/2}\tilde{H}_{l_g-1}^{(k)}W_{l_g}^{(k)}, \tag{3.4}$$

where $\tilde{H}_{l_g}^{(k)} \in \mathbb{R}^{T\times C_{\tilde{H}}\times N}$ denotes the output features for the k-th selected graph after l_g layers, $C_{\tilde{H}}$ is its channel number, A_k denotes the adjacency matrix of the k-th selected graph, Λ_k denotes the degree matrix, normalizing $\hat{G}_{T_k}$ for smooth information propagation. Notably, $W_{l_g}^{(k)}$ refers to the unique set of learnable convolution parameters allocated for the k-th graph, supporting node feature updates according to the specific graph structure. The collective feature set $\tilde{H}_{l_g} \in \mathbb{R}^{T\times C_{\tilde{H}}\times N\times k}$ concatenates the outputs across all selected graphs, creating a comprehensive representation that encapsulates diverse surgical interactions.

Second, our Squeeze-and-Excitation channel attention mechanism [245] facilitates adaptive graph fusion:

$$\begin{aligned} Attn_g &= \sigma(W_g AvgPool(\tilde{H}_{l_g})) \\ H' &= Mean_S(\tilde{H}_{l_g} Attn_g), \end{aligned} \tag{3.5}$$

where $H' \in \mathbb{R}^{T\times C_{H'}\times N}$ represents the attention-weighted geometric graph features, $C_{H'}$ is its feature channel number, $Attn_g \in \mathbb{R}^{T\times S}$ is the attention matrix, σ is the sigmoid activation function, W_g represents the weight parameters, $AvgPool$ denotes the average pooling operation and $Mean_S$ denotes the mean operation along the optimal graph axis.

Third, our l_t-layer dilated causal 2D convolutions model temporal relationships across

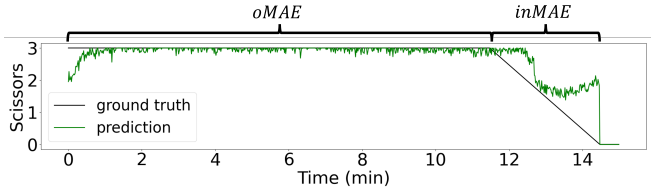

Figure 3.6: Example illustration of evaluating a model for a fixed time horizon of $h = 3\text{min}$. The relevant event (use of scissors) occurs at $t = 14\text{min}$. The ground truth is clipped to be between 0 and h.

frames with more spatio-temporal interaction information where $\hat{H}_0 = H'$:

$$\hat{H}_{l_t} = Conv2D_{l_t}(\hat{H}_{l_t-1}), \tag{3.6}$$

where $\hat{H}_{l_t} \in \mathbb{R}^{C_{\hat{H}} \times T \times N}$ denotes the temporal feature output in l_t layer, $C_{\hat{H}}$ is its channel number, $Conv2D_{l_t}$ denotes causal 2D convolutions that apply 2^{l_t-1} dilation along the time axis to broaden the receptive field in deeper layers, promoting ability to understand broader temporal contexts at a manageable computational cost [21]. The causal design facilitates online inference for each frame [21].

Finally, anticipation outcomes $\hat{Y}_T \in \mathbb{R}^{T \times E}$ are generated from our feature $\hat{H}_{l_t}$ through a 2-layer fully connected neural network with ReLU activation functions.

3.5 Multi-Horizon Objective

A major challenge in previous works [4, 21, 135] is achieving balanced anticipation performance for diverse time horizons. Previous works train and evaluate models using objectives for a fixed time horizon h, typically set at 2, 3, or 5 minutes. The model is supposed to predict the timing in the range of 0 to h, where 0 denotes that the event is currently occurring (*e.g.*, a certain instrument is currently being used) and h denotes that the (beginning of the) event lies h or more in the future (this is also illustrated in Fig. 3.6). A problem with this setup is that setting the horizon h is challenging. When h is small, the model cannot anticipate any long-term events. Conversely, when the horizon is set large, the significant loss for long-term anticipation may compromise the optimization for short-term accuracy.

To address this challenge, we propose a multi-horizon objective training strategy that

enables precise anticipation over different time horizons. Unlike previous methods with different outcomes of separate models for different h, our training and evaluation for different h are based on one model and one outcome $\hat{Y}_T$, which increases the flexibility of our model. We introduce learnable variance representations [246] to evenly distribute objectives across horizons, leading to a more balanced optimization process. This approach enables our training to adapt automatically to the optimal set of horizon weights for various surgical scenarios, enhancing the model's applicability and effectiveness.

In particular, learnable variance representations, denoted by λ, are adopted to adaptively normalize the loss for each training epoch for different tasks (*i.e.,* horizons or different surgical events) [246]. We employ two modified Mean Absolute Errors (MAEs) [4] to form our loss function, optimizing the anticipation for different parts of the ground truth: $inMAE$, where the ground truth is between $(0, h)$; $oMAE$, where the ground truth falls outside the $(0, h)$ range. Illustrations of these can be found in Fig. 3.6. If the event is more than h in the future (out-of-horizon) the model is supposed to return a value of h, otherwise (in-horizon), it is supposed to predict the actual timing. For evaluation, the MAE can be split into its in-horizon part ($inMAE$) and its out-of-horizon part ($oMAE$) [4]: $inMAE$ measures the accuracy of predicting imminent events, while $oMAE$ assesses the ability of a model to recognize that the event is not expected to occur in the near future. The multi-horizon loss L is defined as:

$$\begin{aligned} \hat{\lambda}_h &= \log(1 + e^{\lambda_h}) \\ L &= oMAE_H + \sum_{h}^{H}\left(\frac{inMAE_h}{2\hat{\lambda}_h} + \log(\hat{\lambda}_h)\right), \end{aligned} \tag{3.7}$$

where $oMAE_H$ denotes the $oMAE$ for the largest horizon. For anticipation tasks without a fixed largest horizon, such as remaining surgery duration anticipation, this term is omitted. $inMAE_h$ and $\hat{\lambda}_h$ represent MAE and the learnable parameters for anticipation based on the specific horizon h, respectively. The SoftPlus transformation $\log(1 + e^{(\cdot)})$ keeps the variance representation positive and makes it more suitable for modeling uncertainty in multi-task learning scenarios. The terms $\log(\hat{\lambda}_h)$ serve as a regularizer to prevent risks such as overfitting or learning instabilities from extremely large learnable variances [246].

At the beginning of training, each λ_h is set to 1 to reflect an ideal situation where each horizon is balanced naturally [246]. With this initialization, no specific horizon is favored over others, which prevents any pre-existing bias towards a specific horizon in our learning process for different surgeries.

3.6 Experimental Setup

Data Split and Evaluation Procedures

To demonstrate the effectiveness of our system, we test our method for two types of tasks: 1. Surgical instrument and phase anticipation: They predict the countdown time for when an instrument will be used and when a surgical phase will start, respectively. 2. Remaining Surgical Duration (RSD) anticipation: It predicts the countdown time for when surgery will end.

For surgical instruments and phase anticipation, we primarily used the most common metrics from previous benchmarks [4, 21]. Our main metric is $wMAE$, the average value of $inMAE$ and $oMAE$. As explained in Section 3.5, $inMAE$ measures the error for imminent in-horizon events that lie less than h in the future, while $oMAE$ measures the error for out-of-horizon events that lie more than h in the future. Thus, $wMAE$ provides an overall performance measure that balances in-horizon and out-of-horizon performance. Additionally, we use $eMAE$ [21, 73], which measures the very-short-term prediction accuracy for events that lie less than $0.1\,h$ in the future (e.g. 12 sec for a time horizon of $h = 2$ min). For all these metrics, lower values indicate better performance.

For surgical instruments and phase anticipation, we conducted our comparison on the Cholec80 dataset [2]. It includes 80 laparoscopic cholecystectomy videos, ranging from 15 to 90 minutes in duration. Adhering to the same data splitting and testing protocol in the benchmark [2], we downsampled the videos to 1 Frames Per Second (FPS) for model input, allocating 60 videos for training and 20 for testing. We conducted training four times, each with a different random seed. Following previous works [21], the average performance of both the baseline methods and our method was reported for comparison. In addition, the standard deviation was computed and reported only for our proposed method based on multiple training runs to demonstrate its stability. Our task aimed to

anticipate the occurrence of 5 surgical instruments and 6 surgical phases over horizons of 2, 3, or 5 minutes when they are not occurring [21]. To facilitate meaningful outputs under multi-horizon training, we defined a function $A(O)$ for the model outputs O:

$$A(O) = \begin{cases} h & \text{if } O > h \\ O & \text{otherwise} \end{cases} \tag{3.8}$$

For RSD anticipation, we used the same metrics as previous benchmarks for evaluation [134]: the MAE calculated from the start of 2 minutes remaining, 5 minutes remaining, and from the beginning for the entirety of the minutes remaining. Our comparison is conducted on the Cataract101 dataset [162]. It includes 101 cataract surgeries, with durations ranging from 2 to 20 minutes. Adhering to the same data splitting and testing protocol as established in the benchmark [134], we downsampled the videos to 2.5 FPS and employed 6-fold cross-validation. The test results reflect the average performance across the 6 folds [134]. The primary task focused on predicting the RSD for the entire duration of the procedure [134].

We employ online inference during both training and testing. For consistency with previous works, results on the Cholec80 dataset are reported solely with the mean error [21], while those on the Cataract101 dataset are presented with both mean error and standard deviation [134].

Hyperparameters for Networks and Model Training

The method was implemented in PyTorch 1.10 and trained on a Linux server with an Nvidia GeForce GTX 2080 Ti GPU.

For our models applied to the Cholec80 and Cataract101 datasets, we customized the network hyperparameters to optimize performance for distinct tasks. Specifically, for Cholec80, we set both the policy network layers (l_p) and temporal convolution layers (l_t) to 8, expanding the receptive field to leverage more past information for accurate anticipation. For the Cataract101 dataset, specifically for RSD anticipation, we increased l_p and l_t to 11, enhancing the network capability for long-term understanding. For both datasets, we set the number of initial graphs (C) and the number of selections (k) to

10 and 3, respectively. It is based on our statistical analysis showing that the top 10 graphs cover most interactions. $C = 10$ effectively captures the diversity of dynamic surgical interactions without overfitting to extremely rare interactions in the training data. Setting k to 3 allows us to concentrate on the most relevant candidate graphs for each frame, promoting efficient graph selection.

The horizon sets for our multi-horizon objectives are {2, 3, 5, 7} minutes for the Cholec80 dataset and {2, 5, $+\infty$} (indicating an open-ended horizon) for the Cataract101 dataset. These intervals align with the range of single fixed training objectives found in previous works [4, 21, 134]. They are selected to cover a broad spectrum of short to long-term anticipations. During training, the ground truth values are clipped at the maximum horizon value to prevent overestimation. For model evaluation on benchmarks with predefined horizons, our output is adjusted to match the specified horizon h, facilitating compatibility with established evaluation standards.

We optimized additional hyperparameters using Bayesian hyper-parameter search on the Weights & Biases platform [247]. For the Cholec80 dataset, the optimization determined an epoch count of 100, a learning rate of 0.002, a weight decay of 0.00002, and a batch size of 2. Similarly, for Cataract101, the settings were an epoch count of 100, a learning rate of 0.0003, a weight decay of 0.00005, and a batch size of 4. Furthermore, to facilitate a fair comparison across benchmarks [134], we incorporated auxiliary tasks, including surgical phase recognition and surgeon experience classification, during training to enhance RSD anticipation performance.

Benchmark Models

For the instrument and phase anticipation, we compared our framework with the following recent methods: 1.TimeLSTM [136]: A method that utilizes LSTM to model visual features extracted from videos and requires phase recognition labels during training. 2. RSDNet [133]: A method similar to TimeLSTM [136], but does not require additional annotations. 3. TempAgg [128]: A self-attention-based video anticipation method. 4. B-CNN-LSTM [4]: An initial method for surgical instrument anticipation that uses LSTM to model visual features from videos. 5. IIA-Net [21]: The SOTA method that integrates visual and non-visual features for surgical workflow anticipation. 6. GCN-MSTCN [135]:

Our earlier method proposed in [135], which is an initial approach that uses geometric features as the primary input.

For the RSD anticipation, we compared our framework with the following recent methods: 1) TimeLSTM [136]. 2) RSDNet [133]. 3) TempAgg [128]. 4) CataNet [134]: The SOTA method in this task, which uses a similar design to TimeLSTM [136] and requires phase recognition and surgeon experience labels during training. 5) GCN-MSTCN [135]. We omitted other RSD anticipation works [248, 249] from our comparison due to their different experimental settings.

We sourced performance data directly from SOTA papers [21, 134]. For evaluating TimeLSTM [136], RSDNet [133] on the instrument and phase anticipation task as well as TempAgg [128] for both tasks, we retrained their model using our protocols and metrics. GCN-MSTCN [135], due to the different evaluation metrics used in our earlier attempt, we retrained the model using our protocols and metrics.

3.7 Experimental Results

3.7.1 Comparison with Benchmarks

Table 3.1: $wMAE$ comparison on Cholec80 with **best** and second best scores.

$wMAE$ Comparison	Instrument Anticipation				Phase Anticipation			
	2 min	3 min	5 min	$\text{Mean}_{2,3,5}$	2 min	3 min	5 min	$\text{Mean}_{2,3,5}$
TimeLSTM [136]	0.51	0.76	1.32	0.86	0.47	0.64	1.07	0.72
RSDNet [133]	0.48	0.73	1.26	0.82	0.43	0.63	1.10	0.72
TempAgg [128]	0.65	0.92	1.47	1.01	0.38	0.50	1.18	0.68
B-CNN-LSTM [4]	0.43	0.66	1.09	0.73	0.39	0.59	0.85	0.61
IIA-Net [21]	**0.38**	0.58	**0.92**	**0.63**	0.36	0.49	**0.68**	**0.51**
GCN-MSTCN [135]	0.48	0.72	1.21	0.80	0.45	0.67	1.06	0.73
Ours	0.39±0.002	**0.57±0.01**	1.03±0.01	0.66	**0.35±0.01**	**0.47±0.01**	0.80±0.01	0.54
w/o AG	0.45±0.08	0.70±0.16	1.33±0.30	0.83	0.45±0.15	0.66±0.25	1.16±0.44	0.76
w/o MHO	0.43±0.02	0.61±0.02	1.00±0.02	0.68	0.37±0.01	0.51±0.01	0.80±0.02	0.56

AG: Adaptive Graph Learning; MHO: Multi-Horizon Objective.

Table 3.1 shows the comparison of our proposed method with existing methods in surgical instruments and phase anticipation based on the widely used $wMAE$ metric. It is important to note that other methods, including the SOTA work IIA-Net in Table 3.1, trained individual models for each time horizon. In contrast, our approach trained a single model for all time horizons. Despite this, our method still demonstrates comparable or better performance for the 2-minute and 3-minute time horizons.

In instrument anticipation, our model achieved comparable performance to IIA-Net at the 2-minute horizon and outperformed it by 2% at the 3-minute horizon. For phase anticipation, our method surpassed IIA-Net by 3% and 4% at the 2-minute and 3-minute horizons, respectively. These improvements demonstrate the effectiveness of our framework in short-term anticipation, which is particularly valuable for real-time surgical assistance. For example, during a laparoscopic cholecystectomy, accurately predicting when the dissector will be needed allows the surgical team to prepare the instrument in advance [250], reducing delays and enhancing procedural safety.

Although the absolute $wMAE$ differences between our method and IIA-Net are relatively small in some settings (within $\pm$0.03), they are achieved using a single unified model across all time horizons, whereas IIA-Net requires a separately trained model for each horizon. This highlights the efficiency and generalization ability of our approach.

It is worth noting that our model shows a higher $wMAE$ for phase anticipation at the 5-minute horizon (0.80 vs. 0.68), representing a relative performance drop of approximately 18%. While this is a noticeable gap, its impact is mitigated by the clinical context: shorter horizons (e.g., 2 and 3 minutes) are often more critical for intraoperative decisions, as they typically require short-term alerts [251]. Our stronger performance in those timeframes aligns well with the practical demands of robotic-assisted surgery and real-time surgical planning.

Table 3.2: $eMAE$ comparison with **best** and second best scores.

$eMAE$	Instrument Anticipation			Phase Anticipation		
Comparison	2 min	3 min	5 min	2 min	3 min	5 min
TimeLSTM [136]	1.56	2.26	3.62	1.29	1.73	2.60
RSDNet [133]	1.25	1.80	2.61	1.15	1.55	2.40
TempAgg [128]	1.27	1.35	1.85	1.42	1.49	1.09
B-CNN-LSTM [4]	1.12	1.65	2.68	1.02	1.47	1.54
IIA-Net [21]	1.01	1.46	2.14	1.18	1.42	1.09
GCN-MSTCN [135]	**0.87**	1.26	1.90	**0.77**	1.06	1.46
Ours	0.99±0.01	1.21±0.01	1.49±0.04	0.95±0.05	**1.03±0.04**	**1.06±0.03**
w/o AG	**0.87±0.12**	**1.06±0.13**	**1.34±0.14**	0.94±0.12	1.07±0.19	1.18±0.27
w/o MHO	0.99±0.16	1.42±0.24	2.25±0.30	0.88±0.14	1.17±0.19	1.63±0.23

AG: Adaptive Graph Learning; MHO: Multi-Horizon Objective.

For $eMAE$ (Table 3.2), our framework displayed improvements in both instrument and phase anticipation across all horizons. This indicates that our method is well-suited for rapid-response scenarios, providing accurate predictions even in very short time frames. Lower $eMAE$ scores are particularly important for evaluating very short-term

anticipation, which is crucial for RAS where immediate adjustments are necessary. During RAS, the ability to quickly predict the need for specific instruments or actions significantly enhances the responsiveness of the surgical robots [4]. This further reduces response time and improves the coordination between the surgeon and the robots.

In RSD anticipation (Table 3.3), our model significantly outperforms both CataNet [134] and our previously proposed GCN-MSTCN [135], with improvements of 9% and 57% for the 2-minute horizons, and 9% and 68% for the 5-minute horizons, respectively. The substantial accuracy reduction of GCN-MSTCN suggests the our earlier static graph method may not generalize well for various surgeries. Additionally, even when training our method without auxiliary labels, our method exhibited a significantly smaller performance reduction than CataNet [134], which demonstrates the generalization ability of our method. These error reductions for RSD highlight the potential of our proposed framework to not only benefit planning within a single surgery but also improve operation arrangements across different surgery patients. This advantage enhances the hospital's ability to manage operating rooms more effectively [134], which in turn provides more patients the opportunity to receive timely treatment.

Table 3.3: MAE comparison for RSD Anticipation on Cataract101 with **best** and second best: scores.

Method	2 min	5 min	All	$\text{Mean}_{2,5,\text{All}}$
TimeLSTM [136]	1.22±0.32	1.47±0.78	1.66±0.79	1.45
RSDNet [133]	1.23±0.53	1.37±0.83	1.59±0.69	1.40
TempAgg [128]	0.66±0.41	0.88±0.27	1.47±0.80	1.00
CataNet [134]	0.35±0.20	0.64±0.56	**0.99±0.65**	0.66
CataNet (Only RSD) [134]	0.39±0.28	0.76±0.41	1.11±0.62	0.75
GCN-MSTCN [135]	0.74±0.09	1.82±0.23	3.08±1.40	1.88
Ours	**0.32±0.22**	**0.58±0.27**	**0.99±0.55**	**0.63**
Ours (Only RSD)	0.36±0.26	0.65±0.31	1.00±0.52	0.67

3.7.2 Ablation Study

Ablation studies in Table 3.1 and Table 3.2 evaluate the contributions of the adaptive graph learning module and the multi-horizon training objective. While removing the adaptive graph module slightly improves performance in extreme short-term instrument anticipation (e.g., achieving a lower $eMAE$ of 0.87 at the 2-minute anticipation horizon), it results in a noticeable performance drop in longer-term predictions. For instance, 5-min $wMAE$ rises from 0.66 to 0.83, indicating reduced accuracy when forecasting further into

Table 3.4: $inMAE$ comparison with **best** and second best scores.

$inMAE$ Comparison	Instrument Anticipation			Phase Anticipation		
	2 min	3 min	5 min	2 min	3 min	5 min
TimeLSTM [136]	0.86	1.24	1.98	0.68	0.96	1.54
RSDNet [133]	0.73	1.05	1.62	0.65	0.93	1.55
TempAgg [128]	0.87	1.20	1.78	0.71	0.95	1.40
B-CNN-LSTM [4]	0.77	1.17	1.75	0.63	0.86	1.17
IIA-Net [21]	0.66	0.97	1.40	0.62	0.81	1.08
GCN-MSTCN [135]	**0.64**	**0.93**	1.43	**0.61**	0.86	1.28
Ours	0.72	0.96	**1.33**	0.63	**0.77**	**1.06**

Table 3.5: $oMAE$ comparison with **best** and second best scores.

$oMAE$ Comparison	Instrument Anticipation			Phase Anticipation		
	2 min	3 min	5 min	2 min	3 min	5 min
TimeLSTM [136]	0.16	0.29	0.67	0.22	0.32	0.60
RSDNet [133]	0.23	0.42	0.90	0.20	0.33	0.65
TempAgg [128]	0.43	0.64	1.16	**0.06**	**0.06**	0.97
B-CNN-LSTM [4]	0.08	**0.15**	**0.44**	0.15	0.32	0.52
IIA-Net [21]	0.10	0.19	**0.44**	0.10	0.18	**0.28**
GCN-MSTCN [135]	0.33	0.52	0.99	0.29	0.47	0.85
Ours	**0.07**	0.19	0.73	0.07	0.17	0.55

the future. This trade-off suggests that without graph adaptation, the model may overfit to immediate interactions while ignoring how instrument and target relationships evolve across surgical stages. The static graph fails to model these temporal variations effectively. In contrast, our adaptive graph design dynamically selects relevant interactions for each frame, enabling the model to capture both the immediate context and long-term surgical flow more reliably. In addition, the standard deviations of the results, reported for our proposed method, are lower than those of its ablations, demonstrating the stability of the full model and highlighting that the removal of key components tends to increase performance variability.

By combining this adaptive graph representation with the multi-horizon prediction strategy, our full model consistently achieves the best overall performance, demonstrating stronger robustness and generalization across various time horizons and clinical scenarios.

3.7.3 Detailed Performance Analysis for $wMAE$

For further analyzing our anticipation performance for surgical instruments and phases, we also report the results of $inMAE$ (Table 3.4) and $oMAE$ (Table 3.5), which represent in-horizon and out-of-horizon performance, respectively. Their mean corresponds to the $wMAE$ (Table 3.1). Our results show significant advancements, in particular, in instrument anticipation for $inMAE$, we surpass the SOTA by 5% in the 5-minute time

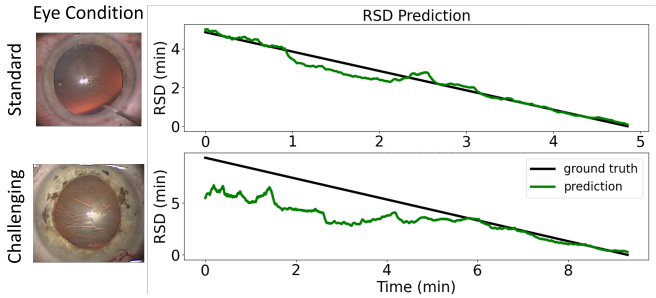

Figure 3.7: RSD anticipation visualization on the Cataract101 dataset. Top: Standard scenario — a standard case without significant inflammation. Bottom: Challenging scenario — a challenging surgery with significant post-inflammation scarring.

horizon, and in phase anticipation, we outperform IIA-Net by 2% in the 5-minute time horizon. Lower $inMAE$ values imply an increased precision in predicting imminent events, ensuring timely interventions during surgery. For example, during cataract surgery, accurate anticipation of the need for specific instruments allows the surgical team to prepare and respond promptly, thereby reducing delays and improving overall surgical efficiency [134]. Additionally, our framework often ranks second-best for $oMAE$. These low $oMAE$ values demonstrate the robustness of our approach in predicting events that occur later in the surgery, which is crucial for long-term procedures where maintaining prediction accuracy over extended periods is essential. This reliability is vital for maintaining operational efficiency and patient safety.

3.7.4 Qualitative Study

Our qualitative study, as illustrated in Fig. 3.7, highlights the adaptability of our system across diverse surgical scenarios. Specifically, on the Cataract101 dataset, we analyzed the performance differences between challenging surgical cases and standard cases. In cataract surgery, significant inflammation in the eye can lead to a challenging case [252]. Hence, we classified our test set, which consists of 20 cases, based on the observation of significant inflammation or post-inflammatory scar tissue. We found that 8 out of the 20 videos had significant inflammation and classified them as challenging cases, while 12 out of 20 were considered normal situations with no significant inflammation. The MAE for these challenging cases is 1.17, while the MAE for the standard cases is 0.88. Since the error in challenging cases is only slightly higher than the overall MAE (including both standard

and challenging cases), this demonstrates that our framework can reliably anticipate both challenging and standard cases. Examples of our anticipation in both standard and challenging cases can be found in Fig. 3.7. Our system performed well with standard cases (Fig. 3.7 Top) but encountered greater errors at the beginning of challenging cases, which often represent more complex surgical procedures. Nonetheless, the reduced countdown speed in difficult cases (Fig. 3.7 Bottom, where dark post-inflammatory scars can be observed around the pupil) indicates our method's ability to adjust its outputs in challenging scenarios.

3.7.5 Robustness Analysis

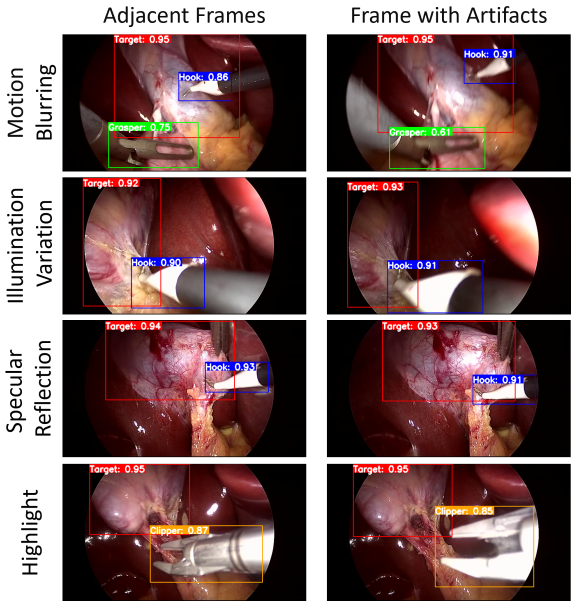

Figure 3.8: Bounding box performance under various common surgical imaging artifacts. Despite the presence of artifacts in frames (**Right**), the position and size of bounding boxes remain consistent with adjacent frames with fewer artifacts (**Left**), demonstrating our method's resilience to varying imaging qualities in surgical endoscopy.

The visualization in Fig. 3.8 illustrates the performance of our bounding box in handling common surgical artifacts, such as motion blurring and illumination variation [46]. Despite the presence of artifacts, the position and size of bounding boxes remain consistent with those in adjacent frames with fewer artifacts. This demonstrates the resilience of our method to varying imaging qualities in surgical videos.

3.7.6 Analysis of Training Horizon Settings

Table 3.6: The effect of different horizon settings for training, $wMAE$ for instrument anticipation. **Bold**:the best scores.

Horizon Setting	2 min	3 min	5 min
2 min	0.53	-	-
2, 3 min	0.40	0.62	-
2, 3, 5 min	0.40	0.59	1.19
2, 3, 5, 7 min	**0.39**	**0.57**	**1.03**
2, 3, 5, 7, 9 min	0.41	0.61	1.12

The impact of different training horizon settings on instrument anticipation is detailed in Table 3.6. The table reveals that extending the training horizon from 2 minutes to 7 minutes consistently improves the $wMAE$ across the 2, 3, and 5-minute anticipations, achieving optimal results of 0.39, 0.57, and 1.03, respectively. Interestingly, further extending the horizon to 9 minutes slightly decreases the performance. Thus, a training horizon of 2, 3, 5, and 7 minutes emerges as the most effective setting. These results demonstrate our assumption that an appropriate multi-horizon training setting offers a global insight for each horizon to augment performance, while also suggesting the efficacy of our adaptive learning strategy.

3.8 Summary

In this chapter, we explored how geometric features serve as robust representations to improve long-term video anticipation in complex surgical settings. Through our investigation, we found that geometric features such as the location and size of surgical instruments and targets, when combined with detection confidence, provide a stable and reliable input even under common visual artifacts like motion blur and illumination variation.

We introduced a novel approach for surgical workflow anticipation using live video data, outperforming existing methods on several benchmarks. Our method extracts geometric features, such as the location and size of surgical instruments and targets, while accounting for detection uncertainty. These features proved particularly robust against common visual artifacts, such as motion blur and variable lighting conditions. We further integrated these features into an adaptive graph learning framework to model

dynamic interactions, enhancing performance. A multi-horizon objective was employed to balance short- and long-term predictions, enabling generalization across varying surgery durations. Notably, our model achieved a 3% improvement in surgical phase anticipation and a 9% improvement in RSD prediction, outperforming specialized state-of-the-art models. By combining geometric features, adaptive graph learning, and a multi-horizon learning strategy, our method improves preparation and coordination within surgical teams, potentially reducing intraoperative risks and enhancing patient outcomes. Additionally, the increased accuracy may optimize operating room resources, enabling more timely treatments for patients.

For future research, we will focus on advancing temporal feature modeling, and developing a more detailed representation of surgical anatomy. First, while our model advances temporal feature modeling, its capacity for long-term dependencies can be enhanced. Future efforts might leverage diffusion-based models to strengthen long-term anticipation [253]. Second, despite the state-of-the-art results of our current model based on bounding boxes, a more detailed representation of surgical anatomy could refine our predictions. Future iterations could integrate wider contextual spatial and visual information to more accurately reflect the surgical dynamics [196].

This chapter demonstrated how geometric features, when combined with adaptive graph learning, could substantially enhance long-term surgical workflow anticipation. While this contribution focuses on temporal prediction, clinical video analysis also demands improvements in the visual quality of input videos themselves [13]. The next chapter (Chapter 4) addresses this by exploring a depth-guided video inpainting method, extending the role of geometric features to the video quality improvement domain.

CHAPTER 4

Endoscopic Video Inpainting Enhanced by Geometric Features

Portions of this chapter have previously been published in the following peer-reviewed publication:

- **Zhang, F. X.**, Chen, S., Xie, X., & Shum, H. P. H. (2024). Depth-aware endoscopic video inpainting. In *Proceedings of the 2024 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '24)*, Springer, Marrakesh, Morocco, pp. 143–153.

In the previous chapter (Chapter 3), we demonstrated how geometric features can strengthen long-term surgical workflow anticipation by improving temporal understanding. However, clinical video analysis also requires high-quality visual inputs to fully realize the benefits of such geometric representations.

Shifting focus from temporal prediction to visual assistance, this chapter explores how geometric features can enhance video quality in clinical settings, particularly through endoscopic video inpainting. While the surgical workflow anticipation task discussed in

Chapter 3 highlighted the importance of robust geometric representations and long-term understanding, endoscopic video inpainting addresses more immediate and practical needs, such as enhancing visual quality. Because pixel information is essential for reconstructing plausible textures in inpainting tasks [140], this chapter emphasizes the fusion of geometric and visual features rather than relying solely on robust geometric representations.

Among various geometric representations, depth information is particularly well suited for inpainting. Although endoscopic cameras lack metric calibration [108], relative depth estimated directly from visual features could sufficiently capture surface continuity and spatial relationships [52]. Instead of measuring real-world distances or delineating precise object boundaries, our approach leverages depth to provide a global geometric context that guides the model in reasoning about occluded regions more coherently.

In this context, precise object boundaries in the depth map are not essential. While high-resolution boundary information is important for tasks such as segmentation or anatomical landmark detection [254], our inpainting framework has different requirements. It uses depth primarily to provide coarse geometric context, including the general layout of anatomical structures, rather than to delineate exact edges. Visual features are already well suited for capturing low-level texture and boundary cues [255]. In contrast, depth maps offer complementary information that enhances spatial reasoning, particularly in cluttered or low-visibility scenes where visual cues may be compromised. By prioritizing relative depth over pixel-accurate delineation, the framework remains robust to estimation noise and occlusions, which are common in endoscopic videos. This design choice improves generalizability across diverse clinical conditions without depending on precise boundary accuracy.

Alternative geometric solutions, such as optical calibration or albedo modeling, can mitigate certain types of corruption by modeling light–camera geometry [256]. However, these methods require specialized hardware setups [257, 258] and are rarely compatible with the standardized equipment used in routine clinical workflows [108]. Moreover, variations in lighting and motion across procedures further hinder their practical deployment [14]. In contrast, inpainting offers a retrospective, hardware-agnostic solution that can be directly applied to monocular endoscopic recordings [13], addressing a wide range

of corruptions such as specular highlights, minor instrument occlusions, and bleeding artifacts [10].

Building on this foundation, this chapter introduces a novel Depth-Aware Endoscopic Video Inpainting (DAEVI) framework, which efficiently fuses geometric and visual features. It incorporates a Spatial-Temporal Guided Depth Estimation module for direct depth estimation from visual features, a Bi-Modal Paired Channel Fusion module for a channel-by-channel fusion of visual and depth information, and a Depth Enhanced Discriminator to assess the fidelity of the RGB-D sequence, comprising the inpainted frames and estimated depth images. Experimental evaluations on established benchmarks demonstrate the superiority of our framework, achieving a 2% improvement in PSNR and a 6% reduction in MSE compared to state-of-the-art methods. Qualitative analyses further validate its ability to inpaint fine details, underscoring the benefits of integrating depth information into endoscopic video inpainting.

4.1 Introduction

In endoscopic videos, occlusions or artifacts, such as reflections or instrument shadows, significantly degrade the visual quality. This issue is commonly known as *corruptions*, hiding critical anatomical details required for endoscopy examinations and surgeries, affecting clinical decisions significantly [259]. The inpainting framework itself does not involve corruption detection [10, 13]. Instead, these corruptions could typically be obtained using highlight-based filters, thresholding methods [260], or directly from existing benchmark datasets that provide annotated masks [13].

As a technique to improve video quality by reconstructing the corrupted regions based on the uncorrupted information, video inpainting is introduced by [10, 13, 46, 145] into the endoscopic scenario to mitigate the corruptions, known as endoscopic video inpainting. While these methods could reconstruct 2D visual information in corrupted endoscopic videos, they suffer from preserving vital 3D spatial details, resulting in artifacts and spatial inconsistency at the inpainted regions, such low-fidelity performance limits their reliability for clinical applications.

Employing depth maps to complement 3D geometric understanding is widely applied

in general video painting [231, 232, 261], which offers a promising solution to preserve 3D spatial awareness for endoscopic video inpainting. Nevertheless, applying this solution is hindered by three significant challenges: First, it is not feasible to pre-acquire endoscopic depth maps, as the depth sensor is not available in standard monocular endoscopic cameras [108]. Second, given the learned depth features from deep-learning-based methods, simply concatenating visual features channel-wise and using vanilla convolution for fusion [261] fails to effectively exploit the depth representation, as they tend to capture redundant representations from visual features [262], losing the 3D spatial details complemented by depth maps. Third, none of these methods [231, 232, 261] assess the 3D spatial fidelity in RGB inpainted outputs, which compromises the reliability of inpainted content.

To address these challenges, we propose the Depth-Aware Endoscopic Video Inpainting (DAEVI) framework that provides more reliable inpainted details for improved clinical reference. It consists of a Spatial-Temporal Guided Depth Estimation (STGDE) module, a Bi-Modal Paired Channel Fusion (BMPCF) module, and a Depth-Enhanced Discriminator (DED), each designed to overcome the respective challenge. First, our STGDE module extracts depth information during visual feature learning to provide 3D spatial information, thus avoiding the requirement for pre-acquired depth maps as input. Second, the BMPCF module conducts a tailor-made feature fusion algorithm to better correlate the 3D spatial relevancy between visual and depth features by pair-wise fusing each visual and depth feature. Third, our DED assesses the 3D spatial fidelity of the RGB-D sequence formed by the inpainted frames and estimated depths, promoting realistic outputs with plausible 3D spatial details.

We evaluate our method on the HyperKvasir endoscopic video dataset [6] and compare it with the corresponding benchmark [13, 61]. The quantitative experiments demonstrate that our proposed DAEVI outperforms state-of-the-art approaches [13], achieving approximately 2% better Peak Signal-to-Noise Ratio (PSNR) and 6% lower Mean Squared Error (MSE). Our qualitative results (Section 4.4.3) show that DAEVI inpaints more fine-grained details, such as microvessels and the boundary of instruments. Furthermore, we directly apply our DAEVI trained on HyperKvasir to the SERV-CT datasets [263], demonstrating our method's generalizability in endoscopic video inpainting.

Our work contributes in several ways, as outlined below:

1. To the best of our knowledge, Depth-Aware Endoscopic Video Inpainting (DAEVI) is the first endoscopic video inpainting framework to incorporate depth information. The effectiveness is demonstrated by comprehensive experiments.

2. We propose a Spatial-Temporal Guided Depth Estimation module to translate depth representation directly from latent visual features, hence circumventing the challenge of acquiring depth maps during endoscopic surgery.

3. We design a Bi-Modal Paired Channel Fusion module that fuses each pair of channels from visual and depth features, which effectively leverages 3D spatial details in endoscopic inpainting.

4. We introduce a Depth-Enhanced Discriminator within our end-to-end optimization, which assesses the fidelity of the inpainted RGB-D sequence, promoting realistic outputs with more plausible 3D spatial details.

4.2 Methods

Given the input endoscopic video frames $X \in \mathbb{R}^{T\times H\times W\times 3}$, we leverage the binary mask $M \in \mathbb{R}^{T\times H\times W\times 1}$, which identifies the corrupted regions, to get the input $X_M = X \odot M$. After processing by our DAEVI to generate the uncorrupted output $\hat{Y} \in \mathbb{R}^{T\times H\times W\times 3}$. $\odot$ is the element-wise product, $H \times W$ is the spatial dimension. The whole formulation is denoted as: $\hat{Y} = DAEVI(X_M)$. This chapter primarily focuses on inpainting rather than corruption detection, so corruption mask generation is not included as part of our framework. The binary masks for corrupted regions can be obtained using highlight-based filters, thresholding methods [260] in live video settings, or directly from existing benchmark datasets that provide annotated masks [13].

The overall architecture is shown in Fig. 4.1. First, DAEVI employs a convolutional encoder to embed X_M into compact visual features $F \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$ to effectively represent local visual features. Following this, our Spatial-Temporal Guided Depth Estimation (STGDE) module learns multi-level visual features and translates them into depth maps. Subsequently, the proposed Bi-Modal Paired Channel Fusion (BMPCF) module fuses

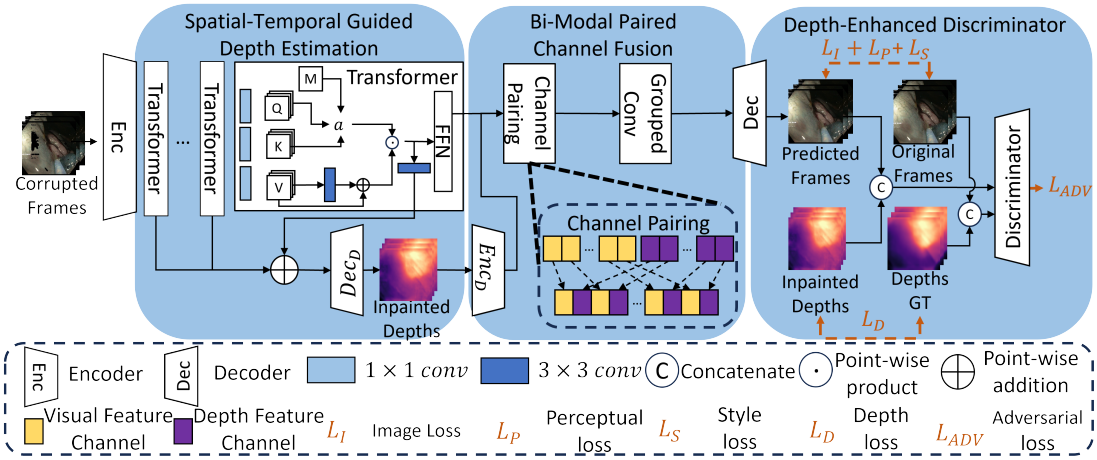

Figure 4.1: The overview of our framework. First, our Spatial-Temporal Guided Depth Estimation module translates depth information from corrupted frames (See 4.2.1). Second, our Bi-Modal Paired Channel Fusion module effectively fuses visual features with depth features (See 4.2.2). Third, our Depth Enhanced Discriminator assesses the fidelity of the inpainted RGB-D sequence (See 4.2.3).

visual and depth features to obtain an integrated representation with enhanced 3D spatial details. After that, a convolutional decoder reconstructs the final inpainted frames $\hat{Y}$. During training, our Depth-Enhanced Discriminator (DED) assesses the visual and spatial fidelities of the inpainted RGB-D sequence.

4.2.1 Spatial-Temporal Guided Depth Estimation (STGDE)

Depth-aware endoscopic video inpainting faces a unique challenge in acquiring depth data, as the standard endoscopic cameras are technically unable to provide the raw depth [108]. To address this challenge, we propose a STGDE module to translate depth features from latent visual features. STGDE involves multiple transformer blocks TB and a depth decoder Dec_D. Each TB is a spatial-temporal transformer block [61] with a spatial enhancement to enhance the local feature learning, thereby improving the representation ability. Dec_D aims to gather the latent visual feature across all TBs for effective depth estimation.

Specifically, after the encoder, corrupted frames X_M are embedded as visual features F, which are fed into the first transformer block TB_1. The output from TB_i is subsequently fed into the following TB_{i+1}, where $i \in (1, N_s - 1)$ and N_s is the number of TB_s. The F^{i-1}, as the input for each TB_i, are linearly transformed into query Q^i, key K^i, and value

V^i separately. Different from STTN [61], we introduce a 3×3 depth-wise convolution $P'^i_V(\cdot)$ to enhance the V^i for better spatial feature learning:

$$\begin{aligned} Q^i &= P^i_Q(F^{i-1}), \\ K^i &= P^i_K(F^{i-1}), \\ V^i &= P^i_V(F^{i-1}) + P'^i_V(F^{i-1}), \end{aligned} \tag{4.1}$$

where $P^i_Q(\cdot)$, $P^i_K(\cdot)$, and $P^i_V(\cdot)$ are 1×1 convolutional layers, and $P'^i_V(\cdot)$ is a 3×3 depth-wise convolution that enhances the spatial sensitivity of V^i. After that, we split each of Q^i, K^i, and V^i into $n = r_1 \times r_2$ smaller patches Q^i_p, K^i_p, and $V^i_p \in \mathbb{R}^{Tn \times c \times h/r_1 \times w/r_2}$, where $h/r_1 \times w/r_2$ is the spatial dimension of patches. Then we utilize the patched Q^i_p, K^i_p, and V^i_p to get the attention output F^i_{att}:

$$S^i = \frac{Q^i_p (K^i_p)^\top}{\sqrt{r_1 \times r_2 \times c}}, \tag{4.2}$$

$$F^i_{att} = \text{softmax}(S^i \odot M) V^i_p, \tag{4.3}$$

where S^i is the attention score, and M denotes a resized binary mask matrix indicating the currupted regions [61]. Then, after a convolutional projection P^i_F followed by a feed-forward network FFN^i, we obtain the output F^i of TB_i:

$$F^i = FFN^i \left(P^i_F \left(F^i_{att} \right) \right). \tag{4.4}$$

To translate depth maps $\hat{D}$ from latent visual features, we aggregate F_{att} across all TB_s to gather the multi-layer visual representation. After a convolutional projection P^i_D, the depth decoder generates the depth maps $\hat{D}$:

$$\hat{D} = Dec_D \left(\sum_{i=1}^{N_s} P^i_D \left(F^i_{att} \right) \right). \tag{4.5}$$

It is important to note that the depth maps generated by the STGDE module are in a relative, rather than metric, scale due to the lack of calibration in standard monocular endoscopic cameras. However, relative depth is sufficient for our framework, as the goal is not to measure absolute distances but to enhance spatial coherence and structural

consistency in the inpainted regions. By capturing spatial dependencies through depth estimation [52], relative depth provides the geometric context necessary to guide more realistic texture reconstruction and enable effective spatial reasoning during inpainting.

4.2.2 Bi-Modal Paired Channel Fusion (BMPCF)

Endoscopic depth-aware inpainting encounters a challenge in effectively integrating depth with visual information, as the simple channel-wise concatenation followed by a vanilla convolution is unable to fully exploit the correlation between depth and visual feature, especially in endoscopic scenes such a complex nature involving varied spatial structures [14].

To address this challenge and effectively enhance visual information with 3D spatial details, we design a BMPCF module to correlate each visual and depth feature by a tailor-made pair-wise fusion algorithm.

Specifically, given the depth maps $\hat{D}$ translated by Dec_D, we first enlarge the channel capacity of $\hat{D}$ with a depth encoder Enc_D to obtain the embedded depth feature F_D, which has the same number of channels as the STGDE's output F^{N_s}. Then, to ensure each depth correlates precisely to the corresponding visual feature, we sequentially interleave the sliced F^{N_s} and F_D in channel-wise:

$$F_{pair}[:, 2i] = \begin{cases} F^{Ns}[:, i] & \text{for } i = 0, 2, \ldots, c-2, \\ F_D[:, i] & \text{for } i = 1, 3, \ldots, c-1, \end{cases} \tag{4.6}$$

where c is the channel number of F^{Ns} and F_D, also indicating the number of pairs. After that, the group-wise convolution $G(\cdot)$ [264] is employed on F_{pair}. This operation divides the channels into groups, where each group comprises one visual channel and one depth channel. Within each group, the convolutional kernel processes the visual and depth features together, facilitating their fusion. The result is the fused output $F_f \in \mathbb{R}^{T\times c\times h\times w}$:

$$F_f = G(F_{pair}). \tag{4.7}$$

In this way, each convolutional kernel facilitates the fusion between every two adjacent channels consisting of one visual feature and one depth feature. Subsequently, a

convolutional decoder reconstructs inpainted frames $\hat{Y}$ from F_f.

4.2.3 Depth-Enhanced Discriminator (DED)

Effectively assessing spatial fidelity is critical in endoscopic depth-aware inpainting, as minor inaccuracies, such as incorrect anatomical details, could significantly impact clinical decisions [265]. While GAN strategies [266] in previous depth-aware inpainting methods [232, 261] only enhance the fidelity of RGB content and ignore the fidelity in 3D spatial details, resulting in unreliable outputs for clinical reference [261]. To this end, we introduce DED during training, to comprehensively assess the RGB-D inpainted endoscopic frames across spatial, temporal, and depth dimensions.

Specifically, we follow [267] to build our DED with 6 3D convolutional blocks to learn both spatial and temporal features for assessment. The inpainted frames and depth are concatenated as the RGB-D data to be the input of DED. To facilitate the assessment of the fidelity of the whole frames after inpainting, both corrupted and non-corrupted regions are included for the adversarial loss $L_{ADV} = L_{GEN} + L_{DED}$ adopted from [266]:

$$\begin{aligned} L_{DED} = \mathbb{E}_{(D,Y)\sim P_{Data}}[Relu(1 - DED([D, Y]))] + \\ \mathbb{E}_{(\hat{D},\hat{Y})\sim P_G}[Relu(DED([\hat{D}, \hat{Y}]))], \end{aligned} \tag{4.8}$$

$$L_{GEN} = -\mathbb{E}_{(\hat{D},\hat{Y})\sim P_G}[DED([\hat{D}, \hat{Y}])], \tag{4.9}$$

To enhance parameter optimization efficiency, we adopt an end-to-end optimization strategy for our DAEVI. Our full loss function is as follows:

$$L = \lambda_D L_D + \lambda_I L_I + \lambda_{GEN} L_{GEN} + \lambda_P L_P + \lambda_S L_S, \tag{4.10}$$

where λ denotes the weight of each loss term. L_D and L_I are the L1 reconstruction loss [143] for translated depth and inpainted frames, respectively. L_P and L_S denote the perceptual loss and style loss, respectively [268]. In each iteration, L and L_{DED} optimize our inpainting network and our DED, respectively.

4.3 Experimental Setting

We evaluate our method against the existing benchmark established [13] on the HyperKvasir endoscopic video dataset [6]. This dataset includes 373 videos with a total of 889,372 frames, 343 for training and 30 for testing. Depth ground truth is derived from a pre-trained endoscopic depth estimator [14] on unmasked frames, which is needed only in training. Following the benchmark setting [13], we employ the same masks and pseudo ground truth from [13] to identify corrupted regions in frames and evaluate Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Mean Squared Error (MSE) specifically for corrupted regions.

We configure the block number $N_s = 8$ and set the weights for λ_D, λ_P, λ_S, λ_I, and λ_{GEN} to [0.1, 0.1, 250, 1, 0.01]. We employ Adam optimizer with learning rate = 1e-4, β_1: 0, β_2: 0.99. All experiments are trained on an NVIDIA TITAN RTX 24G GPU with a batch size of 4 for 200k iterations. Training iterations alternate between selecting 5 random or consecutive frames resized to 288 x 288 pixels. For inference, the model processes every 5 corrupted frames alongside 10 nearby corrupted frames sampled for reference, reassembling them into the full video with a real-time processing speed of approximately 0.03 seconds per frame (≈33.3 FPS).

4.4 Results

4.4.1 Comparison with Existing Methods

In Table 4.1, we benchmark our DAEVI framework against the following methods: 1. Arnold *et al.* [141], a diffusion-based approach. 2. Newson *et al.* [12], which employs temporal patches. 3. Spatial-Temporal Transformer Network (STTN) [61], a transformer-based method for video inpainting. 4. Daher *et al.* [13], the current SOTA method for endoscopic video inpainting, which combines STTN with transfer learning to better adapt transformer attention to the endoscopic domain.

In this comparison, our DAEVI framework achieves a 2% improvement in PSNR and a 6% reduction in MSE compared to the current SOTA method by Daher *et al.*, while also achieving a 1.5% increase in SSIM (0.797 vs. 0.785). These results highlight the significant

Table 4.1: Inpainting Performance Comparison and Ablation Study. w/o STGDE: A pre-trained depth estimator is leveraged for depth estimation instead of STGDE; w/o BMPCF: Simple concatenation is used for fusion instead of BMPCF; w/o DED: A standard RGB discriminator is used in GAN training instead of DED.

Methods	$PSNR_{Crop} \uparrow$	$SSIM_{Crop} \uparrow$	$MSE_{Crop} \downarrow$
Arnold *et al.* [141]	19.909	0.559	895.222
Newson *et al.* [12]	22.27	0.650	543.636
STTN [61]	28.683	<u>0.793</u>	119.541
Daher *et al.* [13]	29.542	0.785	104.719
DAEVI (Full Framework)	**30.126**	**0.797**	**97.873**
w/o STGDE	<u>29.801</u>	0.788	105.150
w/o BMPCF	29.695	0.791	<u>103.861</u>
w/o DED	29.286	**0.797**	108.903

advancements enabled by integrating depth information into endoscopic video inpainting.

Clinically, these improvements are pivotal for enhancing the accuracy and reliability of endoscopic diagnosis. High-quality, inpainted videos with fewer reconstruction errors enable clinicians to visualize critical anatomical structures and pathological features more effectively, even in the presence of obstructions such as smoke, blood, or surgical instruments [10]. For instance, a clear and consistent view of mucosal patterns or vascular structures is essential for identifying lesions, polyps, or tumors during diagnostic endoscopy. Reduced inpainting errors also mitigate the risk of misinterpretation caused by visual artifacts [269], thereby supporting more accurate decision-making in both diagnostic and therapeutic procedures. Moreover, the improved video quality facilitates streamlined workflows by reducing the need for repeated inspections and minimizing overall procedural time [46], ultimately enhancing efficiency and improving patient outcomes.

4.4.2 Ablation Study

Our ablation study in Table 4.1 demonstrates that the complete DAEVI framework achieves the best overall performance, highlighting the importance of each proposed module.

When the STGDE module is replaced with a pre-trained depth estimator, the PSNR decreases from 30.126 to 29.801, and the MSE increases from 97.873 to 105.150. This

result indicates that learning depth features directly from corrupted inputs, as done in STGDE, is more effective than using externally estimated depth maps, which may fail to generalize to masked or occluded areas.

When the BMPCF module is removed and replaced with simple channel-wise concatenation followed by vanilla convolution, the performance further drops. Specifically, the PSNR falls to 29.695 and the MSE rises to 103.861. This suggests that the pair-wise fusion design in BMPCF plays a crucial role in correlating visual and geometric features more effectively, especially in complex endoscopic environments where spatial structure is highly variable.

Finally, when the DED module is replaced with a standard RGB-only discriminator, the PSNR decreases to 29.286, and the MSE increases to 108.903, although the SSIM remains unchanged at 0.797. This outcome suggests that the DED primarily contributes to improving spatial consistency and overall realism, effects that are not fully captured by SSIM. This is because the DED is explicitly designed to enhance 3D spatial fidelity by jointly evaluating RGB and depth coherence during adversarial training. While this improves the structural plausibility and spatial relationships in the inpainted regions, SSIM remains more sensitive to local luminance, contrast, and texture similarity, rather than high-level geometric consistency [270]. As such, the benefits of DED may not always be reflected in SSIM, but are nevertheless crucial for clinically reliable inpainting.

These observations confirm that each component contributes meaningfully to the final performance of the framework, and removing any one of them leads to a measurable reduction in effectiveness.

4.4.3 Qualitative Results

Structural Preservation Ability

To evaluate the structural preservation performance of our inpainting method on corrupted regions, we removed specular and high reflections from the instruments in the HyperKvasir dataset [6] and selected neighboring, less corrupted frames as references. Figure 4.2 illustrates how our method effectively restores fine details, such as microvessels and the interface between instruments and organs, which are crucial for safer and more

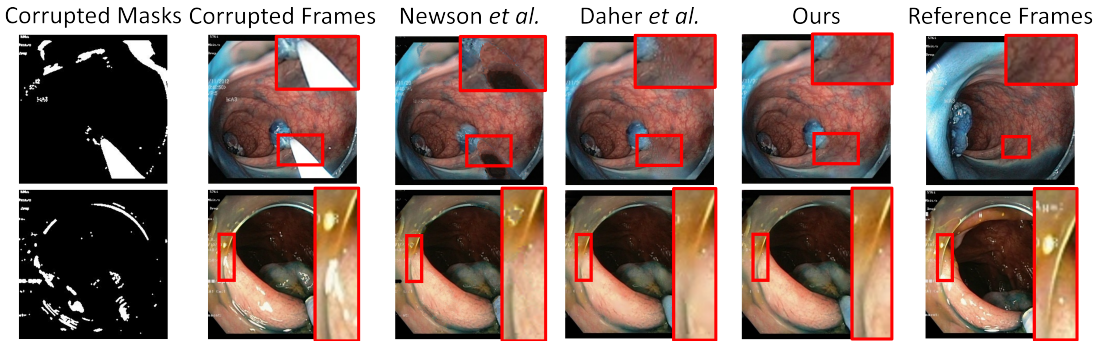

Figure 4.2: Comparison with previous methods by Newson *et al.* [12] and Daher *et al.* [13] on corrupted frames from the HyperKvasir dataset [6]. Red boxes highlight significant differences. Reference frames are near frames with less corruption. Our inpainted content is not only visually plausible but also contextually realistic in terms of spatial structure and temporal consistency.

efficient endoscopic procedures [271].

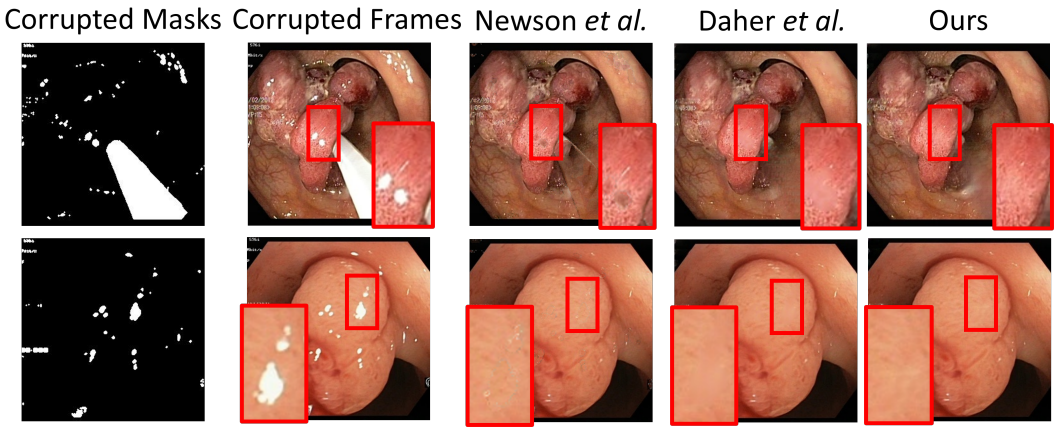

Figure 4.3: More cases from the HyperKvasir dataset [6]: These examples further demonstrate that our method outperforms others, particularly in generating fewer artifacts and more plausible details during endoscopic inpainting. This highlights our approach's superior ability to remove corruption while reconstructing realistic textures.

Texture Reconstruction Ability

Additionally, to assess the performance of our inpainting method in texture reconstruction, we selected other frames for qualitative analysis, as demonstrated in Figure 4.3. These examples highlight how our approach preserves the natural appearance of polyps and other anatomical surfaces in the intestinal tract, excelling in reconstructing realistic textures. Such improvements in texture and detail restoration are essential for accurate diagnosis and successful endoscopic operations.

4.4.4 Generalization Ability Analysis

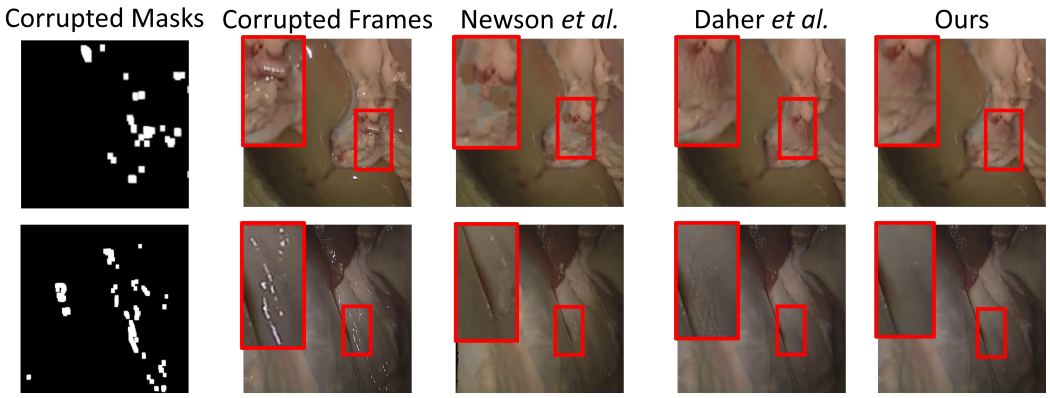

Figure 4.4: Examples from the SERV-CT dataset [6]: These cases demonstrate that our method outperforms others in generating fewer artifacts during inpainting, even without fine-tuning. This highlights the superior generalization capability of our approach.

To evaluate generalization ability, we conducted a zero-shot test on the SERV-CT dataset, an external-body endoscopic dataset with depth ground truth provided [263]. We applied our model to this dataset without any fine-tuning. As shown in Fig. 4.4, our DAEVI method produced the most plausible details in the corrupted regions and generated significantly fewer artifacts than other methods, further demonstrating its superior generalization capability.

4.4.5 Analysis of Depth Preservation Capability

Depth Preservation During Model Inference

We evaluated how well our model reconstructs the spatial structure of input corrupted frames during inference. As shown in Fig. 4.5, the output depth map from our STGDE module (far right) is compared against the depth map estimated by the pre-trained endoscopic depth estimator DepthNet [14] (second from right) on masked corrupted frames. The ground truth depth maps were derived from unmasked frames for reference. Our STGDE module demonstrates more accurate depth estimation, producing depth maps that better align with the ground truth, particularly in preserving the spatial structure of anatomical features. This superior performance highlights the module's ability to reconstruct a more consistent and realistic depth representation, even under challenging conditions. Given that the STGDE output depth map is subsequently used for depth

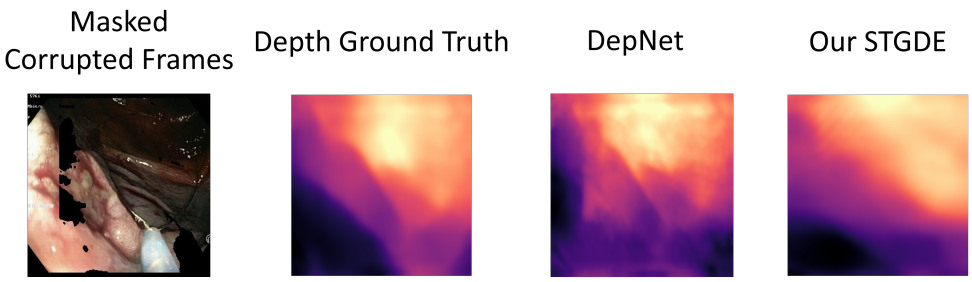

Figure 4.5: Comparison of Depth Estimation Performance for Masked Corrupted Frames: The Spatial-Temporal Guided Depth Estimation (STGDE) module (far right) is compared to a pre-trained endoscopic depth estimator DepthNet [14] (second from right) on masked corrupted frames. Ground truth depth maps derived from unmasked frames (second from left) serve as a reference. The STGDE module demonstrates superior depth estimation, preserving spatial structures and anatomical details more accurately under challenging conditions.

and visual feature fusion in the BMPCF module and adversarial training in the DED module within our DAEVI framework, these results validate that our framework preserves more accurate depth information throughout the model's understanding process. This improved depth preservation directly contributes to the framework's ability to effectively enhance inpainting quality and maintain spatial awareness.

Depth Preservation in Output

We also evaluated the depth preservation capabilities of our inpainted outputs by assessing their retention of spatial details. Figure 4.6 a) demonstrates our method's ability to generate highly plausible content on the SERV-CT dataset without specific fine-tuning, highlighting its robust generalization capabilities. To further assess the preservation of 3D spatial details, we applied a pre-trained DepthNet [14] to the inpainted frames. As shown in Figure 4.6 b), while the visual differences may not be highly pronounced due to the pre-trained DepthNet already being fine-tuned for highlight removal in the previous work [14], our method achieved the lowest Root Mean Squared Error (RMSE) between the ground truth depth and the depth estimations derived from the inpainted frames. These results underscore our framework's effectiveness in preserving depth information and 3D spatial details, outperforming existing methods.

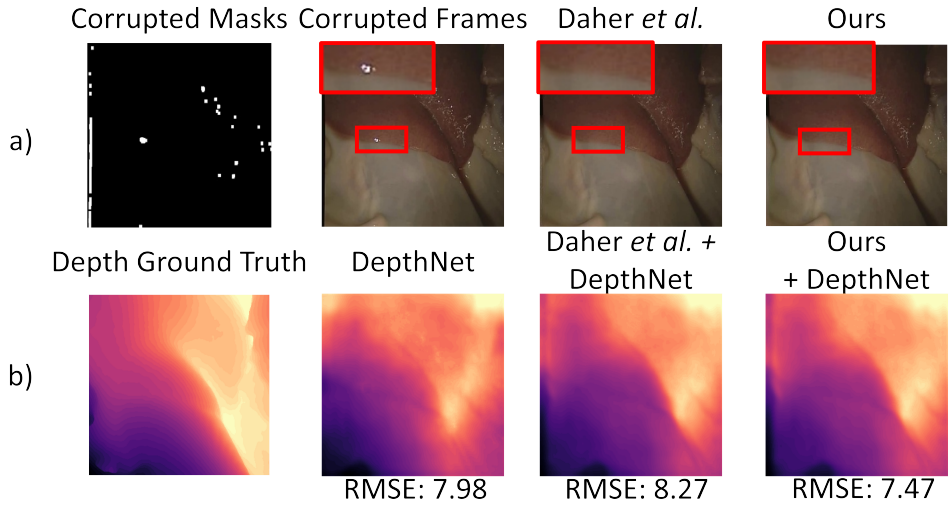

Figure 4.6: Comparison of inpainting performance on the SERV-CT dataset, highlighting depth preservation. (a) Generalization Capability: Inpainted frames generated by our method exhibit higher visual plausibility compared to other methods when tested on the SERV-CT dataset without fine-tuning again. (b) Depth Information Preservation: Using a pre-trained DepthNet [14], we evaluated the accuracy of depth information preserved in the inpainted frames. Our method achieves the lowest RMSE between the estimated depth and ground truth depth maps, demonstrating superior 3D spatial detail retention and outperforming existing approaches.

Table 4.2: Performance Analysis Across Depth Estimation Block Configurations.

Methods	$PSNR_{Crop}\uparrow$	$SSIM_{Crop}\uparrow$	$MSE_{Crop}\downarrow$
DAEVI (First 4 Blocks)	<u>29.921</u>	0.792	<u>100.851</u>
DAEVI (Last 4 Blocks)	29.900	<u>0.796</u>	101.313
DAEVI (All 8 Blocks)	**30.126**	**0.797**	**97.873**

4.4.6 Depth Estimation Block Configuration Analysis

In a deep neural network, shallower layers learn low-level texture features, while deeper layers learn high-level semantics features [272]. Table 4.2 evaluates inpainting performance using different configurations of the first 4, last 4, and all 8 blocks of STGCN for depth estimation within the DAEVI framework. This analysis reveals that employing all 8 blocks, which combines both low-level and high-level features, achieves optimal performance. This demonstrates the efficacy of the STGDE design in integrating multi-level features for optimal inpainting results.

Table 4.3: Online Inference Performance Analysis

Methods	$PSNR_{Crop} \uparrow$	$SSIM_{Crop} \uparrow$	$MSE_{Crop} \downarrow$
Daher *et al.*	29.542	0.785	104.719
DAEVI	**30.126**	**0.797**	**97.873**
DAEVI (Online)	30.117	**0.797**	98.147

4.4.7 Online Inference Performance Analysis

Online real-time inference is essential for immediate endoscopic navigation [273]. However, the existing SOTA method by Daher *et al.* [13] still relies on both past and future frames during inference, falling short of clinical real-time needs. Table 4.3 evaluates the online inference performance of our DAEVI framework by modifying the sampling strategy to use only past frames as reference, making it suitable for real-time applications. Even in this online setting, our framework achieves a PSNR of 30.117, surpassing Daher *et al.* [13], which still uses both past and future frames, by approximately 2%. Similarly, our method maintains a higher SSIM (0.797 vs. 0.785), reflecting a 1.5% improvement in structural similarity. Additionally, we achieve a 6% reduction in MSE compared to Daher *et al.* (98.147 vs. 104.719). These results underscore the robustness of our approach in live endoscopic video processing and highlight its potential for clinical translation, where online real-time performance is a critical requirement.

4.5 Summary

In this chapter, we explore how geometric features can improve clinical video quality through the case of endoscopic video inpainting. Through our investigation, we found that geometric features, particularly depth information, could serve not only as robust representations for future video understanding, as discussed in Chapter 3, but also as powerful priors to directly improve visual recording quality in clinical settings. This demonstrates that geometric features are effective both for anticipating future states and for improving the quality of current visual data, providing immediate support for clinicians.

We propose the DAEVI framework, the first endoscopic video inpainting approach designed to incorporate depth information for achieving reliable 3D spatial details. This

work offers a promising solution to enhance the quality of endoscopic videos, facilitating better clinical decision-making and supporting downstream tasks such as depth estimation [14]. Our DAEVI framework extracts depth information from latent visual features in corrupted endoscopic frames using STGDE, effectively fuses visual and depth information through BMPCF, and assesses the fidelity of the RGB-D content with DED. Quantitative experiments demonstrate that DAEVI outperforms state-of-the-art approaches [13], achieving approximately 2% higher Peak Signal-to-Noise Ratio (PSNR) and 6% lower Mean Squared Error (MSE). These results highlight the significant improvements achieved by integrating depth information. This chapter illustrates how the incorporation of geometric features can enhance clinical video quality, even in highly demanding endoscopic scenarios, while also underscoring the potential for similar advancements in less challenging clinical environments.

While our framework demonstrates improvements in inpainting performance, it is not without limitations. As our work focuses on corruption inpainting, the real-world applicability of DAEVI depends on the quality of the external corruption detection backbone. An inaccurate mask may result in undesired consequences. For example, if diagnostically relevant regions are mistakenly labeled as corrupted, inpainting could unintentionally obscure pathological cues that clinicians rely on. To reduce this risk, future work could incorporate more reliable endoscopic corruption detection techniques, such as semantic segmentation or reflection-aware highlight detection [46], and involve clinical expertise to help distinguish between technical artifacts and diagnostically meaningful content. In clinical settings, the inpainting output should be used as a supportive reference rather than a replacement for direct interpretation of the original video data.

While this chapter focused on improving the visual quality of clinical videos, a deeper understanding of clinical performance also demands fine-grained modeling of human–object interactions across multiple viewpoints. Therefore, the next chapter (Chapter 5) explores structured multi-view modeling, where geometric features such as human skeletons and object relations are leveraged to achieve robust skill assessment in clinical procedures.

CHAPTER 5

Clinical Skills Assessment Enhanced by Multi-view Geometric Features

Portions of this chapter have previously been published or will be submitted to the following peer-reviewed publications:

- Constable, M. D., **Zhang, F. X.**, Conner, T., Monk, D., Rajsic, J., Ford, C., Park, L. J., Platt, A., Porteous, D., Grierson, L., & Shum, H. P. H. (2024). Advancing Healthcare Practice and Education via Data Sharing: Demonstrating the Utility of Open Data by Training an Artificial Intelligence Model to Assess Cardiopulmonary Resuscitation Skills. In *Advances in Health Sciences Education*, pp. 1-21.
- **Zhang, F. X.**, Yao, H., Chen, S., Jia, X., Zheng, S., & Shum, H. P. H. (2025). Towards cross-view multimodality action quality assessment for Traditional Chinese Medicine physical therapy. Rejected with invitation to resubmit in March 2025; revision in preparation for *IEEE Transactions on Instrumentation and Measurement*.

Building on the previous chapters (Chapter 3, Chapter 4) where geometric features enhanced temporal understanding and visual quality, this chapter advances their role

towards a new application: clinical skill assessment. While anticipation and inpainting improve video analysis, they are not sufficient for tasks that demand deeper semantic understanding. Clinical skill assessment presents a more structured and practical challenge, requiring fine-grained semantic interpretation to distinguish subtle variations in motion smoothness, posture alignment, and force control. Unlike video anticipation or quality improvement, skill assessment benefits from standardized procedures and controlled environments, reducing variability and regulatory complexity and making real-world deployment of automated feedback systems more feasible [274]. Here, we explore how multi-view geometric features can bridge the gap between passive video analysis and active performance feedback, enabling structured, objective, and scalable evaluation in medical education.

Clinical skill assessment presents a unique challenge that requires fine-grained semantic understanding. In this context, fine-grained refers to the ability to detect subtle differences in how a task is performed [275]. These include variations in motion smoothness, posture alignment, force control, or spatial accuracy during the same procedure. Rather than focusing on what action is taking place, the task centers on evaluating how well the action is executed [68]. This involves interpreting detailed visual and geometric cues that reflect the performer's technique and level of proficiency. Such distinctions carry important clinical meaning and are necessary for providing meaningful feedback. Single-view systems often miss important information due to occlusions or limited perspective, which limits their ability to capture the depth of performance quality needed for reliable assessment.

To address these challenges, this chapter introduces the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework, which combines geometric and visual features from multiple views for clinical skill assessment. The CME-AQA framework includes an Attention based Visual-Pose Fusion (AVPF) module for feature fusion and an Multiscale View Alignment (MVA) strategy that enables learning from multiple views while still allowing single view inference during deployment.

5.1 Introduction

Clinical skill assessment, as an application of Action Quality Assessment (AQA) in clinical settings, employs various computer vision methods to evaluate the quality of actions performed during therapy sessions. It has proven effective in providing feedback for sports and rehabilitation training [276, 277]. Predominantly, these frameworks utilize pose-based data captured using sensor-based or markerless methods to detail body positions and movements, which are crucial for minimizing environmental noise and capturing detailed motion information [278]. However, they often analyze full-body motions from a single-view perspective and ignore interactions with the environment. This limitation is particularly significant in many common physical therapy practices, such as Traditional Chinese Medicine (TCM) physical therapy, including acupuncture and tui na, where training emphasizes the precision of hand motions frequently obscured due to self-occlusion [279]. Additionally, varied interactions between hands and objects, such as acupuncture needles or practice pads, highlight the critical need to incorporate enhanced environmental visual information [280].

To address these limitations in current frameworks for clinical skill assessment, we have identified two major challenges. First, existing approaches, particularly in physical therapy training [22, 158, 159], rely on pose estimation data that fail to capture the environmental context, including the interactions between hands and various objects and surfaces during therapy sessions. This limitation restricts the effectiveness of clinical skill assessment methods in evaluating all relevant aspects of physical therapy. Second, current skill assessment methods [281, 282], typically based on single-view video settings, do not adequately address frequent self-occlusions, where parts of the body often obscure other parts. This shortcoming compromises the reliability of these methods in real-world applications.

To overcome these challenges in applying existing clinical skill methods, we propose a Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework for clinical skill assessment. This framework integrates view translation of both pose and visual modality features between two different viewpoints. It consists of two main components: the Attention based Visual-Pose Fusion (AVPF) module and Multiscale View Alignment (MVA) training strategy, each designed to tackle specific challenges. First, the

AVPF module utilizes cross-attention to employ visuals as the query and pose data as the key and value, fusing the most relevant information between visual and pose features to provide a comprehensive feature representation for the practice video. Second, the MVA training strategy leverages different view videos in training, allowing our framework to have multi-view awareness even when provided with only a single view input video.

To evaluate the effectiveness of the proposed CME-AQA framework across different types of clinical skills, we selected two representative applications: Traditional Chinese Medicine (TCM) procedures and Cardiopulmonary Resuscitation (CPR). These tasks offer distinct challenges and help assess the framework's ability to generalize across diverse clinical competencies. TCM procedures such as acupuncture and tui na involve complex hand techniques and localized manipulations [283, 284], including precise control over needle depth, angle, and movement frequency. These actions require accurate motor coordination and spatial awareness, making them a rigorous test case for evaluating the model's ability to capture clinically meaningful motion patterns. In contrast, CPR emphasizes coordinated full-body movements that must maintain consistent force and rhythm [285], requiring robust recognition of postural alignment and temporal regularity. These contrasting requirements allow for a comprehensive assessment of our framework's generalizability across diverse clinical competencies.

To facilitate experimentation and broaden the dataset resources available for future research, we introduce two novel multi-view video datasets for TCM and CPR skill assessment. Experimental results demonstrate that CME-AQA significantly enhances the evaluation of complex clinical actions, achieving over 30% higher accuracy and F1 scores in tasks such as Needle Depth and Quick Needle Movements. On the CPR dataset, our framework performs comparably to human experts. These impressive results verify the assumption of this chapter: integrating multi-view geometric and visual information enables more accurate and reliable assessment of complex clinical skills. By capturing both global body posture and fine-grained hand-object interactions across different tasks, the proposed framework demonstrates strong potential for real-world deployment in diverse medical training contexts, paving the way for scalable, data-driven skill evaluation in clinical education.

The source code and the annotated dataset, including sample videos and labels, used

in our work are available on GitHub[1]. Our main contributions are as follows:

1. To the best of our knowledge, we propose the first Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework for clinical skill assessment that integrates both pose and visual modality features. Additionally, we have created TCM-AQA61, a new multi-view video dataset containing first-person and third-person videos of 61 subjects practicing two common TCM physical therapy treatments: acupuncture and Chinese massage.

2. We introduce the Attention based Visual-Pose Fusion (AVPF) module that efficiently fuses the most relevant information between visual and pose features to provide a comprehensive feature representation for the practice video, thereby enhancing the environmental description.

3. We design a Multiscale View Alignment (MVA) training strategy that leverages both egocentric and exocentric view videos in training, allowing our framework to have multi-view awareness even when provided with only a single view input video.

4. We additionally created a multi-view video dataset for CPR and compared our proposed method with human experts to demonstrate its real-world translation potential.

5.2 Data Collection

To the best of our knowledge, no public dataset for multi-view clinical training was available, prompting us to create the TCM-AQA61 dataset for this chapter. This dataset features simultaneous dual-view recordings, as illustrated in Fig. 5.1: a first-person view (*i.e.,* egocentric view) and a third-person view (*i.e.,* exocentric view). We focused on TCM physical therapy, specifically acupuncture and tui na, due to their prevalence, their demonstrated potential in treating various motion disorders [286], and the complex hand interactions they involve [283], which present significant challenges for existing clinical

[1] https://github.com/FrancisXZhang/cme-aqa

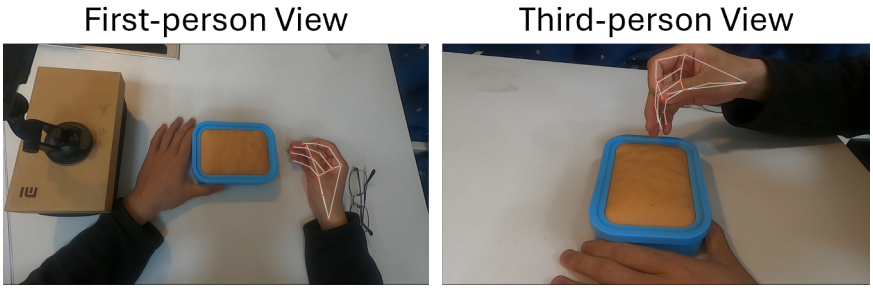

Figure 5.1: Example frame from TCM-AQA61. Video recordings were captured using two GoPro 8 motion cameras: one mounted on the subject's forehead to provide a first-person perspective, and another positioned to capture a side view of the subject's hands. This setup was designed to simultaneously capture hand motions from both the practitioner's perspective and an observer's perspective, which could provide comprehensive insights into hand-object interactions.

skill assessment methods. The dual-view setup was specifically designed to capture hand motions from both the practitioner's perspective and an observer's perspective, providing comprehensive insights into hand-object interactions and enabling the verification of the effectiveness of leveraging cross-view geometric features.

In detail, the data collection encompassed simulated acupuncture and tui na practice sessions conducted by 61 medical students from the Beijing University of Chinese Medicine. Instead of practicing on real patients, participants performed their techniques on silicone pads designed for training purposes. These pads provide a safe and standardized surface for simulating needle insertion and massage maneuvers. Video recordings were captured using two GoPro 8 cameras: one mounted on the subject's forehead for an egocentric view and one positioned laterally to capture hand motions. The cameras were time-synchronized but not geometrically calibrated, as our method does not rely on multi-view triangulation or 3D pose lifting. Instead, feature alignment is handled at the representation level through our multi-view training strategy, which is robust to viewpoint variations and does not require precise camera geometry. We chose a head-mounted egocentric camera over a fixed front-facing one to capture fine hand-object interactions from the practitioner's perspective [287]. This dynamic viewpoint enhances visibility of subtle actions often occluded in static views and reflects the visual focus of the operator, which is critical for assessing clinically relevant skills. It also supports future extensions into VR-based immersive training [288].

After data collection, two experienced TCM physical therapists (HY and SC) conducted

two rounds of manual AQA for each subject. Videos from both views were available to them during their review. In the first round, the therapists reviewed the videos individually. In the second round, they discussed their differing ratings to reach a consensus for each skill assessed. The skill subjects for acupuncture included Needle Holding, Needle Angle, Needle Depth, Quick Needle Movements, Lifting and Thrusting Frequency, Lifting and Thrusting Amplitude, Twisting Frequency, Twisting Amplitude, and Quick Needle Withdrawal. For tuina, the assessed skills included Sinking Shoulders, dropping elbows, Suspended Wrists, Hollow Palms, Solid Fingers, Elbow/Forearm Force, Depth, and Frequency. Notably, although the absolute needle depth is difficult to measure from video alone, the needle length was fixed during data collection. This allowed the therapists to estimate insertion depth by observing the relative change in needle insertion across consecutive frames and comparing the visible remaining length of the needle. Based on these observations, they were able to assess and annotate whether the needle depth was clinically appropriate.

Sample videos along with their corresponding labels are available in our GitHub repository[2].

5.3 Methods

5.3.1 Methodology Overview

Our overview figure is shown in Fig. 5.2. During inference, the input consists of both visual and geometric features extracted from a raw video captured from a single third-person view, which is the most commonly used perspective in clinical skill assessment [155]. Visual features are extracted using a pre-trained ResNet [181], while 3D hand skeletons (x, y, and relative z) are obtained using MediaPipe Hands [220], a two-stage hand tracking model that first detects hand regions with a palm detector and then regresses 21 hand landmarks with relative depth using a neural network. Although the resulting 3D coordinates are not metrically calibrated, they provide sufficient geometric context for clinical skill analysis. These extracted features serve as compact and informative inputs

[2] https://github.com/FrancisXZhang/cme-aqa

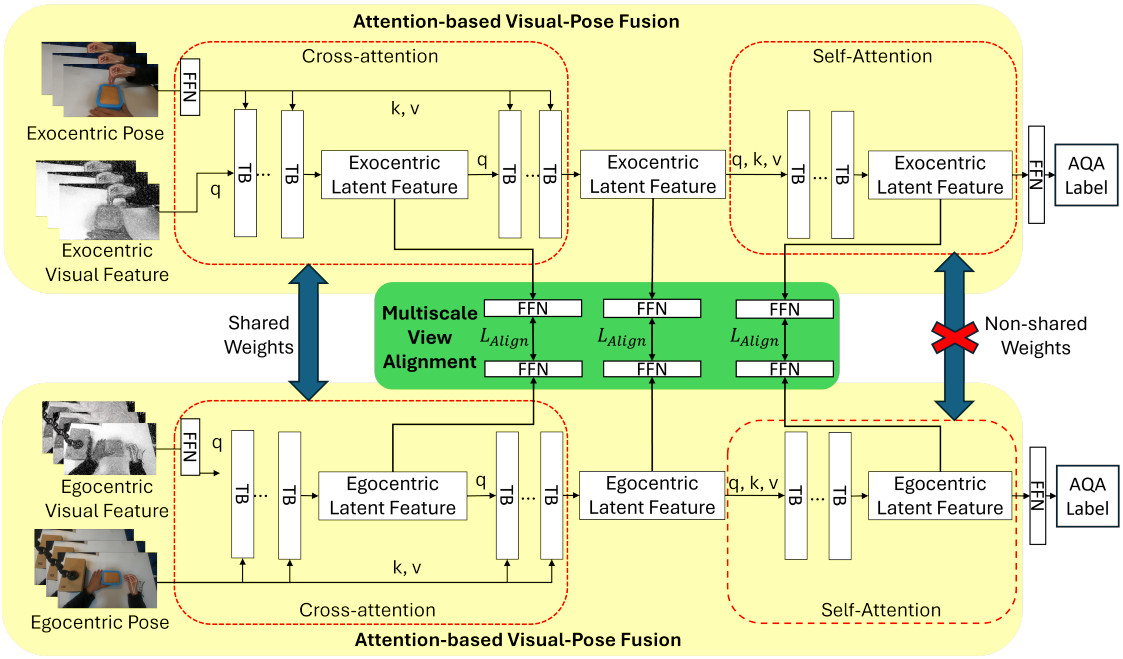

Figure 5.2: Overview of the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework when applied to the TCM-AQA61 dataset, where **TB** refers to the transformer block. The framework comprises two main components designed to leverage both first-person view (*i.e.,* egocentric view) and a third-person view (*i.e.,* exocentric view) to train an AQA framework for TCM physical therapy: (1) The Attention based Visual-Pose Fusion (AVPF) module (see Section 5.3.2), which employs a cross-attention mechanism to fuse visual and pose features, enhancing the environmental description of the practice video by correlating the most relevant visual and pose features. (2) The Multiscale View Alignment (MVA) training strategy (see Section 5.3.3), which aligns features across different scales from the AVPF module to maintain invariant features between egocentric and exocentric views, thereby enabling a more comprehensive feature representation with awareness of both views. Notably, the presentation of this overview figure and methodology emphasizes first-person and third-person views to align with the TCM-AQA61 dataset. However, this setup may not be required for other clinical practices, as demonstrated in Section 5.5, where our method generalizes effectively to CPR training with a front-view and side-view configuration.

for the learnable components of our framework.

The cross-attention mechanism in the Attention based Visual-Pose Fusion (AVPF) module first correlates the visual and pose features, enhancing the integration of these modalities and enabling more efficient encoding of dynamic hand-object interactions. The self-attention mechanism then models temporal dependencies in the sequence, and finally, fully connected layers predict scores for multiple clinically relevant skill items.

During training, the framework uses visual and geometric features from both first-person and third-person views via the Multiscale View Alignment (MVA) strategy. This design allows the model to align multi-view features across different scales and learn more robust, view-invariant representations. At inference time, only third-person view input is needed, making the system more practical and lightweight while retaining the benefits of multi-view supervision.

It is worth noting that the presentation of the overview figure and methodology is designed to clearly illustrate our method alongside the TCM-AQA61 dataset, with a focus on the use of both first-person and third-person views. However, when applying our framework to other clinical practices, this dual-view setup may not be necessary. As demonstrated in Section 5.5, our method generalizes effectively to CPR training, even with a front-view and side-view configuration.

5.3.2 Attention based Visual-Pose Fusion (AVPF) Module

A major challenge in applying existing methods [22,159,235] is their tendency to disregard visual information, relying solely on human skeleton data. This approach is inadequate for complex practices such as acupuncture [284], where complex movements involving the needle, hand, and practice pad are critical. Relying exclusively on human skeleton data fails to capture sufficient visual detail, thereby limiting the accuracy of inferences in clinical training scenarios involving dense hand-object interactions.

To address this challenge, our AVPF module employs a multi-layer cross-attention mechanism [289] to efficiently fuse the most relevant information between visual and pose features. This approach not only provides a comprehensive feature representation with environmental context, but also progressively enriches the visual features with detailed pose information, thereby maintaining a structural understanding of human

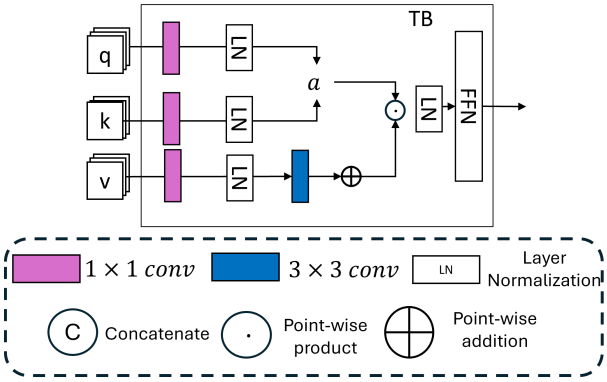

Figure 5.3: The architecture of our layer-normalization enhanced attention transformer block. To improve robustness against feature inconsistencies caused by occlusions or lighting variations in complex clinical settings [15], layer normalization is applied after the 1D convolution to stabilize the modeling of the attention matrix.

skeleton data. Unlike direct fusion by concatenation, which fuses global features from different modalities [290], this cross-attention method allows the module to focus on relevant features across modalities and reduce redundancy, thereby improving the overall accuracy of the AQA process.

In detail, visual features are extracted using an off-the-shelf, pre-trained visual feature extraction model to create a compact feature representation, denoted as $F_V \in \mathbb{R}^{T\times C}$, where T represents the video length and C represents the number of feature channels. Similarly, pose features are extracted by a pre-trained 3D pose estimation pipeline to describe the movements of the practitioner in 3D space. Following extraction, the human skeleton pose features, denoted as $F_P \in \mathbb{R}^{T\times C}$, are further processed by a multi-layer Feedforward Neural Network (FFN) to adjust the pose feature dimensions to match those of F_V, thereby enhancing the efficiency of the pose feature representation. Each cross-attention transformer block, indexed by i, processes F_V^i and F_P through our specially designed layer-norm-enhanced attention mechanism (see Fig. 5.3) to mitigate noise commonly appearing in visual and pose feature representations due to occlusion or varying lighting conditions in complex practice scenarios [15]:

$$\text{Attn}^i(Q,K,V) = \text{softmax}\left(\frac{LN(Q)LN(K)^T}{\sqrt{d_k}}\right)LN(V), \tag{5.1}$$

where Q is the query matrix derived from F_V^i via 1D convolution, and K and V are the key and value matrices derived from F_P via 1D convolution. LN denotes the layer

normalization used to smooth the feature representation. After the attention computation, a multi-layer FFN embeds the attention scores into the updated F_V^i for the next layer, continuing until the ith layer to produce the final fused feature F_V^I. This selective updating of only F_V allows us to iteratively refine the visual features with pose data while maintaining a stable pose reference across updates, promoting greater consistency and reliability in action modeling.

Then, the refined fused feature F_V^I from our cross-attention mechanism then serves as the input to our designed layer-norm-enhanced self-attention mechanism, which is similar in design to that illustrated in Fig. 5.3. Each self-attention transformer block, indexed by j, processes the features:

$$\text{Attn}^j(Q, K, V) = \text{softmax}\left(\frac{LN(Q)LN(K)^T}{\sqrt{d_k}}\right) LN(V), \tag{5.2}$$

where Q is the query matrix and K and V are the key and value matrices derived from F_V^j via 1D convolutions. The process continues until the jth transformer layer to produce F_V^J.

Finally, F_V^J is input into a two-layer fully connected network to translate the latent features into predicted action quality scores for each clinically relevant item.

5.3.3 Multiscale View Alignment (MVA) Training Strategy

Another challenge in applying common AQA methods to assess complex clinical practices is that existing methods [22, 159, 235] typically rely solely on a single third-person view for both training and inference. However, in many clinical practices, such as acupuncture, crucial hand actions often suffer from self-occlusion when recorded from a single perspective [291].

To address this challenge while maintaining a simple deployment process, our MVA training strategy incorporates two different views during training, utilizing both egocentric and exocentric perspectives, but requires only a single-view input during inference. This strategy aligns latent features at multiple stages, including the cross-attention process, the output features of cross-attention, and the self-attention outputs of the AVPF module. By aligning features across views at multiple scales, from early feature modeling

to higher-level abstractions, the framework is better equipped to learn invariant features at various hierarchical levels [292]. This integration of dual-view features enhances the framework's capacity to capture detailed semantics for skill assessment while allowing single-view inference to benefit from multi-view training. In detail, the MVA alignment loss is calculated as follows:

$$L_{\text{Align}} = \sum_{m=1}^{n} \lambda_m \cdot \|F_{V,\text{View1}}^m - F_{V,\text{View2}}^m\|_1, \tag{5.3}$$

where $F_{V,\text{View1}}^m$ and $F_{V,\text{View2}}^m$ represent the feature representations at the m-th scale from two different views (*e.g.,* first-person view and third-person view, respectively, when training on the TCM-AQA61 dataset), $\|\cdot\|_1$ represents the L1 norm which measures the distance between these two feature sets, and λ_m represent the weights assigned to each scale, emphasizing the importance of alignment at different levels of feature abstraction. This calculation aims to minimize discrepancies between different views, potentially enhancing the model's ability to generalize from both viewpoints.

The overall training loss for the framework is defined as:

$$L = \alpha L_{\text{Align}} + \beta L_{\text{AQA}}, \tag{5.4}$$

where L_{Align} represents the cumulative L1 norm losses between corresponding features in different views, and L_{AQA} is the cross-entropy loss between the predicted quality assessment labels $\hat{Y}_{\text{AQA}}$ and the true labels Y_{AQA}. The parameters α and β are weighting factors that balance the importance of alignment and AQA accuracy in the overall loss function. During training, the cross-attention in AVPF employs shared weights between the different views to capture invariant correlations between visual and pose features. Conversely, the self-attention in AVPF employs non-shared weights to ensure that the final feature representation is specifically tailored and adaptive to the AQA tasks.

5.4 Experiment Design

We split our dataset into 5 folds to conduct cross-validation. We configured our network hyperparameters with $J = 8$, $L = 4$, $\alpha = 0.5$, and $\beta = 0.5$. The batch size is set at 2,

and the training duration is 100 epochs. We employ the Adam optimizer with a learning rate of 1×10^{-4}, $\beta_1 = 0$, and $\beta_2 = 0.99$. All experiments were conducted on an NVIDIA TITAN RTX 24G GPU.

For each skill subject, we comprehensively evaluated our approach using two metrics: accuracy and F1 score, both expressed as unitless proportions ranging from 0 to 1. This metric selection is justified because our task aligns more closely with other clinical skill assessment tasks for clinical applications, which are often regarded as multiple-category classification [68]. Accuracy measures the proportion of correctly identified actions, reflecting the overall system performance. The F1 score balances precision (the proportion of correctly identified actions among those predicted) and recall (the proportion of correct actions that were actually identified). Clinically, high accuracy and F1 scores indicate that the system provides precise and comprehensive feedback, which is crucial for effective TCM physical therapy training.

5.4.1 Performance Metrics Analysis

We compare our proposed methods with the following baselines: 1. STGCN [234]: A classical pose-based human action recognition method using graph neural networks. 2. STNN [22]: An early method that leverages human pose for assessing physical rehabilitation exercises via Long Short-Term Memory (LSTM) [97]. 3. STGCN-LSTM [158]: A method that extends STGCN with LSTM to assess physical rehabilitation exercises using human pose data. 4. STGCN-RI [235]: A method that leverages human pose and joint rotation matrices to assess physical rehabilitation exercises via STGCN. For fair comparison, all baseline methods are retrained on our proposed dataset.

The results presented in Tables 5.1 and 5.2 demonstrate that our proposed CME-AQA framework consistently outperforms other baseline methods across various skill subjects in both acupuncture and tuina tasks. In acupuncture tasks, our framework achieves the highest average accuracy/F1 score (0.83/0.71) across skills, compared to 0.75/0.54 for the next best baseline, STGCN-LSTM [158]. Significant improvements are observed in tasks requiring fine-grained motion understanding, such as Needle Depth and Quick Needle Movements, where our framework exceeds baseline accuracy by over 30% relative improvement. Similarly, in Twisting Frequency, our method outperforms

Table 5.1: Performance Comparison of Acupuncture Skills (Measured by Accuracy/F1 Score)

	Needle Holding	Needle Angle	Needle Depth	Quick Needle Movements	Lifting & Thrusting Frequency	Lifting & Thrusting Amplitude	Twisting Frequency	Twisting Amplitude	Quick Needle Withdrawal	Average Across Skills
STGCN [234]	0.73/0.42	0.80/0.74	0.62/0.59	0.42/0.40	0.78/**0.67**	0.61/0.51	0.85/0.62	0.80/0.49	0.75/0.49	0.70/0.54
STNN [22]	0.90/0.57	0.68/0.40	0.59/0.45	0.41/0.28	0.75/0.42	0.65/0.49	0.83/0.45	0.93/0.68	0.87/0.56	0.73/0.47
STGCN-LSTM [158]	0.90/0.57	0.72/0.57	0.65/0.61	0.42/0.41	0.78/0.56	0.65/0.51	0.83/0.45	0.93/0.68	0.87/0.56	0.75/0.54
STGCN-RI [235]	0.76/0.46	0.52/0.51	0.63/0.59	0.52/0.49	0.76/0.57	0.47/0.42	0.72/0.53	0.93/**0.78**	0.87/0.61	0.68/0.55
Ours	**0.91/0.62**	**0.84/0.78**	**0.87/0.87**	**0.70/0.67**	**0.80/0.67**	**0.69/0.62**	**0.90/0.81**	**0.95**/0.75	**0.88/0.66**	**0.83/0.71**

Table 5.2: Performance comparison for tuina skills (Measured by Accuracy/F1 Score)

	Standard Action	Hollow Palm	Solid Fingers	Slow Movement	Depth	Frequency	Average Across Skills
STGCN [234]	0.68/0.63	0.70/0.60	0.93/0.68	**0.98/0.89**	0.91/0.57	0.66/0.64	0.81/0.66
STNN [22]	0.72/0.55	0.73/0.53	**0.94**/0.68	**0.98/0.89**	0.95/0.68	0.44/0.36	0.79/0.61
STGCN-LSTM [158]	0.70/0.59	0.75/0.56	**0.94**/0.68	**0.98/0.89**	0.95/0.68	0.64/0.55	0.82/0.65
STGCN-RI [235]	0.74/0.72	0.68/0.62	0.93/**0.79**	0.94/0.68	0.95/0.68	**0.85/0.84**	0.84/0.72
Ours	**0.83/0.81**	**0.80/0.74**	**0.94**/0.68	**0.98/0.89**	**0.98/0.89**	0.81/0.79	**0.89/0.80**

Table 5.3: Ablation Study of Acupuncture Skills (Measured by Accuracy/F1 Score)

		Needle Holding	Needle Angle	Needle Depth	Quick Needle Movements	Lifting & Thrusting Frequency	Lifting & Thrusting Amplitude	Twisting Frequency	Twisting Amplitude	Quick Needle Withdrawal	Average Across Skills
Modality & View	Ours	**0.91**/0.62	**0.84/0.78**	**0.87/0.87**	0.70/0.67	0.80/0.67	0.69/0.62	**0.90/0.81**	**0.95/0.75**	**0.88/0.66**	**0.83/0.71**
	(w/o TPV)	**0.91**/0.62	0.73/0.69	0.71/0.65	0.74/0.72	0.81/0.67	0.67/0.57	0.85/0.62	0.93/0.68	**0.88/0.66**	0.80/0.65
	(w/o FPV)	**0.91**/0.65	0.80/0.77	0.78/0.78	0.79/0.78	0.87/**0.80**	0.69/0.63	0.83/0.45	0.95/0.75	0.87/0.56	**0.83**/0.68
	(w/o Visual)	0.90/**0.68**	0.72/0.67	0.83/0.83	0.77/0.75	**0.88/0.80**	**0.74/0.67**	0.88/0.69	0.93/0.68	0.85/0.65	**0.83/0.71**
	(w/o Pose)	0.90/0.57	0.78/0.76	0.76/0.74	**0.79/0.77**	0.78/0.63	0.72/0.66	0.83/0.45	0.93/0.68	0.87/0.56	0.81/0.64
Design	Ours	**0.91/0.62**	**0.84**/0.78	**0.87/0.87**	0.70/0.67	**0.80/0.67**	0.69/0.62	**0.90/0.81**	**0.95/0.75**	**0.88/0.66**	**0.83/0.71**
	(w/o AVPF)	0.90/0.57	0.77/0.68	0.67/0.64	0.72/0.69	0.76/0.47	0.61/0.51	0.85/0.52	0.93/0.68	0.87/0.56	0.78/0.59
	(w/o MVA)	**0.91/0.62**	0.83/**0.79**	0.71/0.68	**0.84/0.83**	0.77/0.59	**0.79/0.69**	0.83/0.45	**0.95/0.75**	**0.88/0.66**	**0.83**/0.67
Share Weight	Ours	**0.91/0.62**	**0.84/0.78**	0.87/**0.87**	0.70/0.67	0.80/0.67	0.69/0.62	**0.90/0.81**	**0.95/0.75**	**0.88/0.66**	0.83/**0.71**
	Late	0.90/0.57	0.79/0.75	**0.88**/0.87	0.76/0.75	0.80/0.69	0.67/0.60	0.83/0.50	0.93/0.68	0.87/0.51	0.82/0.65
	No share	**0.91/0.62**	0.69/0.59	0.82/0.81	0.67/0.66	**0.90/0.79**	0.74/0.63	0.85/0.57	0.91/0.67	**0.88/0.66**	0.81/0.66
	All share	0.90/0.57	0.80/0.74	0.82/0.81	**0.82/0.81**	0.81/0.72	**0.79/0.68**	0.83/0.45	0.93/0.68	0.87/0.56	**0.84**/0.66

previous approaches by over 10% in accuracy.

In tuina tasks, our method also demonstrates superior performance, achieving the best average accuracy/F1 score (0.89/0.80) compared to 0.84/0.72 for STGCN-RI [235]. Notably, our method achieves over 15% higher F1 scores in Standard Action and Hollow Palm compared to STNN [22] and STGCN-RI, while maintaining robust performance in depth and frequency estimation.

This superior performance can be attributed to the effective integration of third-person view insights and egocentric features, which enhance the model's ability to capture fine-grained details and reduce the effects of occlusion. By leveraging these multi-view perspectives, the CME-AQA framework demonstrates strong generalization and applicability across diverse clinical training scenarios.

Regarding real-world applicability, the accuracy values in Table 5.1 range from 0.70 to 0.95, reflecting the proportion of correctly classified skill assessments. In procedures such as acupuncture and tui na, small differences in technique and hand-object interaction are important [293] and are often difficult for human evaluators to assess consistently [294]. Achieving accuracy above 0.70 across multiple skill categories suggests that the system is capable of reliably distinguishing between different performance levels. The high F1 scores, reaching up to 0.87, indicate a strong balance between precision and recall. This balance is essential for providing consistent and meaningful feedback in clinical training.

5.4.2 Ablation Study

To comprehensively validate the effectiveness of each component in our proposed framework, we conducted a detailed ablation study covering three aspects: view and modality, framework design, and weight sharing strategies, as summarized in Table 5.3.

Overall, our complete model achieves the highest and most balanced performance, with an average accuracy of 0.83 and F1 score of 0.71 across all acupuncture skills. This highlights the effectiveness of integrating multi-view information, multimodal features, and specialized architectural designs. Beyond the average scores, our full model also demonstrates consistent robustness across diverse skills, especially in tasks requiring fine-grained motion understanding, such as Twisting Frequency and Needle Depth.

View and Modality Ablation

We assessed the impact of removing specific views (first-person view or third-person view) and modalities (visual features or pose features) to understand their importance. The results demonstrate that integrating both views and modalities leads to the best performance. Notably, removing the first-person view caused significant drops in accuracy and F1 scores for skills requiring fine hand movements, such as twisting and lifting, underscoring the value of multiple viewpoints in mitigating occlusion issues. Similarly, omitting pose features notably affected tasks that rely on precise spatial awareness such as needle depth, indicating that both visual and pose features are essential and contribute uniquely to the overall assessment.

Framework Design Ablation

We assessed the impact of removing key components of the framework. For the configuration without AVPF, we replaced the cross-attention fusion mechanism with simple fully connected layers, resulting in a significant performance drop, particularly in tasks requiring complex action interpretation such as Needle Angle. This highlights the effectiveness of cross-attention in fusing pose and visual data. Additionally, for the configuration without MVA, where only the final output was aligned (as opposed to multiscale alignment), we observed decreased performance across most metrics, particularly in tasks requiring temporal consistency, such as Twisting Frequency. This suggests that multiscale feature alignment plays a critical role in improving feature learning at different hierarchical levels.

Weight Sharing Design

We explored various weight-sharing strategies within the model. In the "late share" configuration, we did not share weights for cross-attention but shared weights for self-attention in the AVPF module. In the "no share" configuration, no weights were shared between different view inputs. Lastly, in the "all share" configuration, weights were shared for both cross-attention and self-attention across views. The results indicate that our current design where weights are shared for cross-attention but not for self-attention, achieved the best performance. This suggests that while cross-attention benefits from

shared weights to capture invariant correlations across views, self-attention requires view-specific adaptations to achieve more accurate feature representation and improved performance.

5.5 Generalized Analysis for CPR Skill Assessment

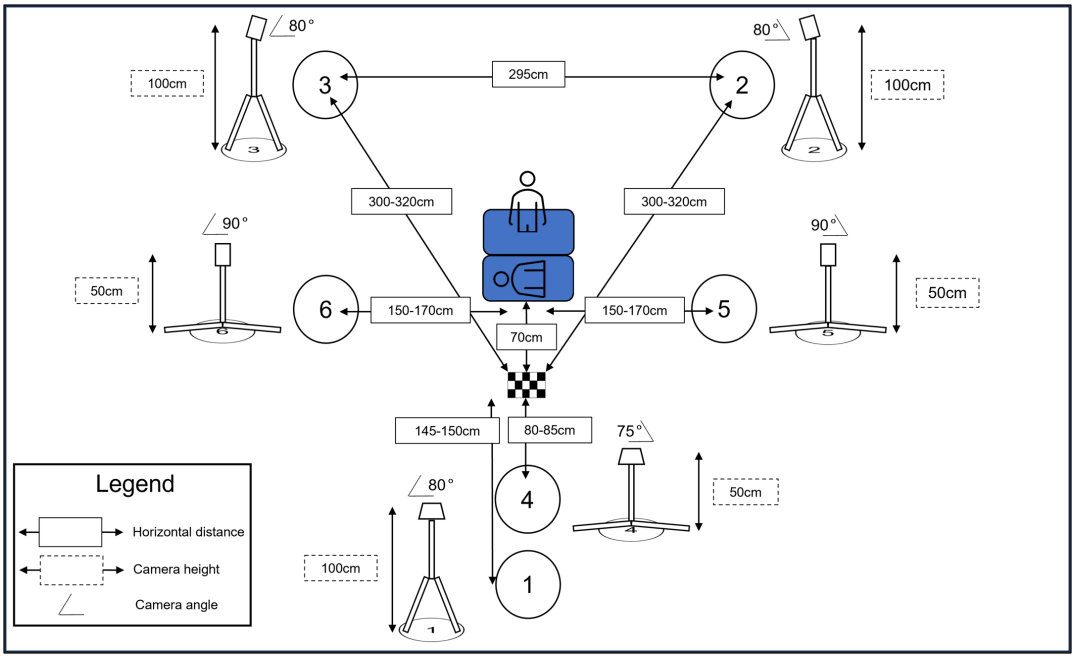

Figure 5.4: The recording setup for our CPR dataset, including camera positions and the task area. Circles depict the location of the cameras. The checkerboard was placed in front of the task space with one foam mat for the manikin and one foam mat for the participant. Approximate distances between cameras are provided, although there was some slight variation between days.

To validate the generalizability of our method across different practices and view configurations, as well as contribute to the development of multi-view clinical training datasets, we further constructed a multi-view dataset for CPR training and compared our model's performance to that of human experts. When applying our CME-AQA framework to this dataset, we utilized a front-view and side-view configuration, which differs from the first-person and third-person views used in the TCM-AQA61 dataset. Additionally, due to ethical agreements, substantial portions of the raw videos in this CPR dataset were blurred to protect privacy, limiting the extraction of meaningful visual features. Consequently, only skeleton data was employed in the CME-AQA framework for this

analysis.

Data Collection

We collected a multi-view CPR training dataset from 53 participants, comprising nursing students and staff from Northumbria University's Department of Nursing and Midwifery. Each participant performed four cycles of 30 chest compressions on a manikin, with recordings captured using six synchronized GoPro cameras strategically positioned to provide comprehensive coverage (see Fig. 5.4). Performance evaluations focused on key metrics [285], including hand position, arm posture, shoulder position, compression depth, compression release, and rate. Each of these items was scored from 0 to 4 by expert raters, reflecting the number of compression cycles that met the expected criteria. These scores indicate the count of successfully performed cycles rather than a physical measurement. Raters independently evaluated each cycle for each item and assigned one point for correct execution, with any disagreements resolved through consensus to ensure consistency and alignment with clinical standards. The study was conducted in compliance with ethical standards approved by Northumbria's Ethics System (No. 44602).

Evaluation Setting

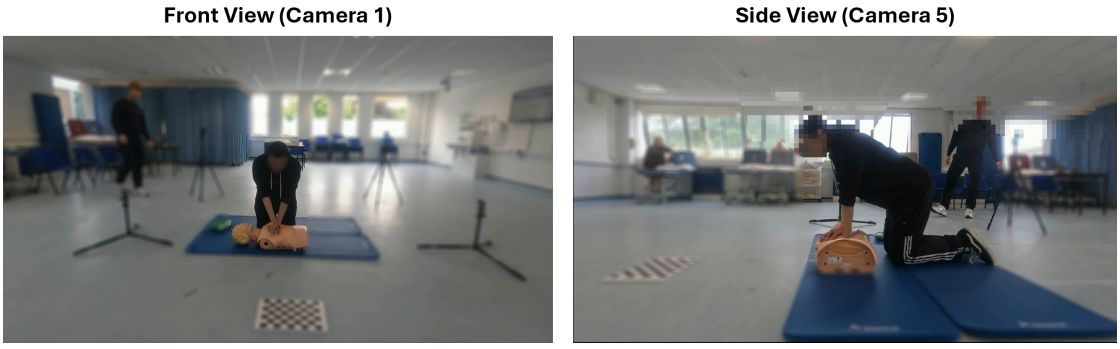

Figure 5.5: Example frames for the views used during training. To facilitate fair comparisons, data from the front view (camera 1) and the side view (camera 5) were used for training, as these were the primary viewpoints utilized by experts during their evaluations. Notably, only the front view was used for inference.

The evaluation of our automatic AQA framework employed Mean Absolute Error (MAE) with fivefold cross-validation. In each fold, 80% of the data was allocated for training and 20% for testing, with training epochs set to 100.

Unlike our TCM application, which involved a single practice session naturally suited for classification, the CPR skill evaluation used cumulative scores ranging from 0 to 4 for each item. These scores reflect the number of correctly performed compression cycles across four CPR rounds and represent varying degrees of performance rather than binary outcomes such as "correct" or "incorrect." Although the labels are discrete, we framed this as a regression task because the scores are ordinal and evenly spaced, allowing the model to learn the relative severity of mistakes. This approach captures detailed differences in performance better than treating each score as an independent class. To evaluate performance, we used Mean Absolute Error (MAE), which has the same units as the original measurement (*i.e.,* the rating score in our CPR assessment) and provides a direct measure of average prediction error. MAE was appropriate for our evaluation, as it clearly indicates how much the predicted scores differ from the true ratings. Since the expert consensus scores served as the ground truth for MAE calculation, the MAE quantified the average error between the model predictions and the consensus scores. Similarly, it quantified the average error between individual expert ratings and the consensus.

To facilitate fair comparisons, data from the front view (camera 1) and the side view (camera 5) were used during training, as these were the primary viewpoints utilized by experts during their evaluations. Notably, only the front view was used for inference. Example frames of these two views are shown in Fig. 5.5.

Results

Table 5.4: Mean Average Error (Human Experts vs. Our CME-AQA Framework)

Item	Evaluator 1	Evaluator 2	MV-STGCN [295]	STGCN-RI [235]	Ours
Hand Position	1.62	1.08	**0.33**	0.39	**0.33**
Arm Position	0.70	0.15	**0.07**	0.13	**0.07**
Shoulder Position	0.40	0.34	**0.13**	0.19	**0.13**
Depth	0.49	**0.30**	0.69	0.77	0.69
Rate	0.89	**0.11**	1.67	1.81	1.81
Compression Release	1.04	**0.98**	1.00	1.06	1.02
Total	3.96	**2.69**	2.98	3.14	2.96

As shown in Table 5.4, Evaluator 1 and Evaluator 2 represent the MAE between the experts' individual ratings and the final consensus scores. Our CME-AQA framework achieved a total MAE of 2.96, calculated by comparing the sum of ratings for each subject

with the sum of expert consensus scores. This performance closely aligns with that of the human evaluators (3.96 for Evaluator 1 and 2.69 for Evaluator 2). Compared to the baseline method, STGCN-RI [235], which achieved a total MAE of 3.14, our framework demonstrates improved accuracy across most metrics.

Notably, our framework achieved lower MAE scores in critical metrics such as Hand Position (0.34) and Arm Position (0.08), outperforming both human evaluators and the baseline method. Slight discrepancies were observed in metrics such as Depth and Rate, likely due to the frequent interactions between the practitioner and the manikin in these tasks. Since only human skeleton data was used as input in this analysis due to data limitations, capturing these interactions fully proved challenging. This further supports our assumption that visual information is also crucial in clinical skill assessment. Nevertheless, our framework maintained accuracy comparable to both human evaluators and the baseline method in these categories. These results highlight the potential of our CME-AQA framework to provide reliable automated feedback during clinical training sessions, supporting its viability for real-world clinical applications.

5.6 Summary

In this chapter, we explored how multi-view geometric features can enhance fine-grained semantic understanding, using clinical skill assessment as the validation task. Through our study, we found that multi-view geometric features, when combined with visual information, significantly improve the ability to capture subtle performance differences in clinical skills, enabling more accurate evaluation. This builds on previous chapters: compared to Chapter 3, where geometric features were used to improve temporal understanding, and Chapter 4, where they guided video inpainting, here geometric features are extended toward fine-grained semantic interpretation by leveraging multi-view perspectives. By moving beyond single-view limitations, we enable a more comprehensive and clinically meaningful understanding of practitioner movements.

The main contribution of this chapter is the introduction of the Cross-view Multi-modality Enhanced Action Quality Assessment (CME-AQA) framework, which incorporates the Attention based Visual-Pose Fusion (AVPF) module for efficient fusion of

visual and pose features and the Multiscale View Alignment (MVA) training strategy for multi-view awareness. Notably, this framework maintains its effectiveness even with single-view input during inference. Additionally, this chapter introduces two multi-view video datasets: one for TCM physical therapy and another for CPR. Our evaluations demonstrated that integrating multi-view geometric features significantly improves clinical skill assessment accuracy. Specifically, CME-AQA enhanced assessment performance, particularly in complex tasks such as Needle Depth and Quick Needle Movements, achieving over 30% higher accuracy and F1 scores. Furthermore, in experiments with the CPR dataset, CME-AQA delivered performance comparable to human experts. This comparable performance underscores its strong potential for real-world translation.

Although this work has shown promising results for fine-grained semantic understanding of clinical videos, particularly in differentiating performance levels for the same clinical technique, several avenues for future research remain. First, our framework relied on existing pose estimation methods to extract human skeleton data as geometric features. These methods, often lacking temporal awareness, can produce inconsistent or noisy outputs [296], potentially impacting downstream assessments. Future efforts could focus on integrating pose estimation into an end-to-end learning process that incorporates temporal information, thereby enhancing estimation accuracy. Second, to evaluate the real-world impact of our framework, future research could deploy an offline version of the system to junior clinical students and monitor its effectiveness in improving their skills over time. This would provide valuable insights into the practical benefits of multi-view geometric features for clinical skill development.

CHAPTER 6

Conclusion

In this thesis, we systematically explored the evolving role of geometric feature enhanced deep learning in clinical video analysis. Starting from structured geometric inputs for modeling complex workflows, advancing through the fusion of geometric and visual features for enhancing clinical video quality, and finally extending to multi-view geometric feature integration for assessment of practitioner skills, we demonstrated how geometric representations can drive robust, real-time, and fine-grained insights that surpass traditional methods.

Our findings explicitly demonstrate that geometric features are not merely auxiliary signals but fundamental enablers for clinical video analysis. Whether structured as primary inputs, fused with visual modalities, or integrated across multiple views, geometric representations consistently enhanced robustness and real-world applicability across diverse clinical tasks. This thesis thus establishes geometric feature enhanced deep learning as a central strategy for advancing reliability, scalability, and clinical utility in clinical video analysis.

6.1 Achievement of Aims and Objectives

This thesis represents a deliberate and progressive expansion of the role of geometric features in clinical video analysis: starting from structured inputs for understanding dynamic processes, advancing through the fusion of geometric and visual features for enhancing clinical video quality, and leveraging multi-view geometric features for fine-grained skill evaluation.

The primary aim of this thesis was to explore how geometric feature enhanced deep learning can address the key challenges in clinical video analysis. The progress made in each technical objective demonstrates how this approach systematically evolved to meet increasingly complex demands, as detailed below.

In Chapter 3, we laid the foundation by addressing the objective of enhancing real-time long-term anticipation. We proposed a robust geometric feature representation that includes instruments, surgical targets, and their detection confidence values. This structured input enabled accurate modeling of dynamic workflows and surgical interactions. The dynamic graph selection method further allowed the adaptive representation of diverse procedural contexts. These advancements demonstrated the utility of geometric features as primary structured inputs for high-level predictive tasks, establishing a foundation for future collaborative and semi-autonomous surgical systems.

Building upon this foundation, Chapter 4 addressed the challenge of improving video quality in extreme clinical environments. Here, we extended the role of geometric features by fusing them with visual modalities in the DAEVI framework. This fusion achieved depth-aware reconstruction and enhanced video quality, even under severe degradation. By preserving anatomical details and spatial structures, this work demonstrated how geometric features could not only structure understanding but also directly enhance visual clarity, meeting real-time diagnostic and interventional needs.

Finally, Chapter 5 expanded the scope further by introducing multi-view geometric feature integration for fine-grained semantic understanding. The CME-AQA framework combined visual and pose features from multiple viewpoints to mitigate self-occlusion and capture comprehensive practitioner movements. By creating and releasing the TCM-AQA61 and CPR datasets, we enabled broader research opportunities in clinical training. This final step moved beyond immediate clinical needs to support educational settings,

highlighting the versatility and scalability of geometric feature enhanced learning in controlled environments.

In summary, the integration of geometric features into deep learning frameworks has demonstrated varying degrees of real-world applicability across different objectives:

- **High-level Teamwork and Robotic System Assistance:** The use of geometric features as the primary input supports anticipation tasks, offering high-level assistance in surgical team dynamics and robotic systems.
- **Direct Visual Assistance for Clinicians:** The fusion of geometric and visual features provides enhanced visual outputs that directly aid clinicians in diagnostic and interventional tasks.
- **Transferable Solutions for Clinical Training:** The leveraging of multi-view geometric features ensures broad applicability in non-critical settings like clinical training, where safety constraints are less stringent, allowing for rapid deployment and feedback in real-world scenarios.

Each chapter thus represents a step in a coherent progression: from structured modeling of dynamic clinical processes, to enhancing visual realism, to achieving fine-grained semantic understanding across views. Ultimately, this journey demonstrates the potential of geometric feature enhanced deep learning to address diverse and critical demands in clinical video analysis.

6.2 Review of Contributions

The main contributions of this thesis are summarized as follows:

- We demonstrated that using geometric features as primary inputs enhances long-term temporal understanding in clinical video analysis, particularly for anticipation tasks. To this end, we proposed an adaptive graph learning framework for surgical workflow anticipation that incorporates both surgical instruments and target anatomy. The framework uses bounding boxes to extract geometric features related to instruments and targets, integrating their detection confidence levels. Our

method dynamically selects candidate graphs to represent interactions between surgical instruments and targets for each timeframe, leveraging graph and temporal convolutions to effectively utilize dynamic geometric features and improve predictions in complex clinical settings (see Chapter 3).

- We demonstrated that fusing geometric features and visual features enhances the quality of clinical videos, particularly for video inpainting tasks. To achieve this, we introduced a novel endoscopic video inpainting framework that integrates depth information to achieve reliable 3D spatial details. This framework extracts depth information during visual feature learning, eliminating the need for pre-acquired depth maps. It employs a tailored feature fusion algorithm that correlates 3D spatial relevance by pairwise fusing each visual feature with its corresponding depth feature. Additionally, it assesses the 3D spatial fidelity of the RGB-D sequence formed by the inpainted frames and estimated depths, enabling realistic outputs with plausible 3D spatial details (see Chapter 4).
- We demonstrated that the introduction of multi-view geometric features enhances the semantic understanding of clinical videos, particularly in clinical skill assessment. To achieve this, we proposed a novel framework that fuses features from multiple views, utilizing graph convolution on skeleton data derived from pose estimation to facilitate precise performance assessments. The framework integrates visual and geometric features while transferring cross-view features to different views, enhancing robustness to frequent hand movements. To validate this approach, we collected and released a dataset for TCM physical therapy assessment, annotated by two clinical experts. Additionally, we created and released a comprehensive dataset of multi-view CPR training videos, complete with skill ratings provided by two clinical experts, and demonstrated that our proposed method achieves performance comparable to human experts (see Chapter 5).

6.3 Future Research Directions

The integration of geometric features with deep learning in clinical video analysis has demonstrated substantial potential, but several areas remain unexplored and ripe for

future research to enhance model robustness, clinical applicability, and scalability.

6.3.1 Adaptive Geometric Features Selection

Further research could explore adaptive geometric feature selection in dynamic clinical environments, enabling models to identify and prioritize the most relevant features for specific tasks automatically. Current methods [21] often rely on prior assumptions about the importance of certain geometric representations, which may not generalize across diverse clinical scenarios. For instance, in routine surgical procedures with clear anatomical visibility, keypoint detection may capture critical landmarks effectively [297]. However, for minimally invasive procedures with close instrument-anatomy interactions, bounding boxes may be more suitable for tracking movements and spatial relationships [132]. Addressing these variations requires a flexible framework that dynamically adjusts feature selection based on the task. A promising solution involves meta-learning combined with model pruning [298], allowing models to evaluate multiple geometric features, such as keypoints, bounding boxes, and depth maps, and retain only the most relevant ones. This approach optimizes computational efficiency, adapts to varying conditions, and enhances model performance in clinical video analysis. However, its implementation demands substantial data and computational resources, highlighting the need for efficient and scalable methods [299].

6.3.2 Privacy-Preserving Clinical Video Analysis

While geometric features such as key points and bounding boxes inherently offer a level of privacy protection by abstracting spatial information without exposing sensitive details, integrating them with visual features can raise privacy concerns, especially in clinical settings where patient confidentiality is critical [300]. This challenge also complicates data sharing across clinical centers [301], limiting the availability of sufficient data for deep learning model training. A promising solution is Federated Learning (FL) [301], which enables institutions to train models collaboratively while keeping data local. However, applying FL to clinical video analysis is challenging due to the non-Independently and Identically Distributed (IID) nature of video data, characterized by intricate temporal

dependencies that complicate splits for validation across institutions [302]. Additionally, the limited size of clinical video datasets often biases model evaluations [302], reflecting dataset-specific traits rather than true generalization. Addressing these issues is crucial for leveraging federated learning in clinical video analysis while preserving privacy and ensuring robust, generalizable performance.

6.3.3 Causality-Driven Explainable Deep Learning

Explainable deep learning is essential in clinical video analysis to build trust among patients and clinicians [303], facilitating the acceptance of deep learning-assisted systems. While geometric features such as key points offer a potentially interpretable means of explaining model inferences, especially for tasks such as clinical diagnosis [304], existing explainability methods primarily focus on highlighting weights or attention values within model layers. For example, Zhang et al. [305] visualize the attention values of human joints in their framework for Parkinson's tremor detection, demonstrating which joints the model prioritizes to make inferences. However, in complex clinical tasks such as surgical workflow anticipation [21], understanding the importance of individual points alone is insufficient. While these points may reveal correlations, temporal causality provides deeper insights by clarifying the sequence and interaction of events that lead to specific outcomes [306], which is crucial in capturing the dynamics of long-term video analysis. Future work could focus on developing causality-based explanations in clinical video analysis, providing clearer, more actionable reports for clinical end-users to foster trust and confidence in deep learning systems.

6.3.4 Human-in-the-Loop Clinical Video Analysis

Another concern for deep learning in clinical video analysis is ensuring safety when the model operates autonomously. Incorporating clinicians into the model development and evaluation process, often referred to as *human-in-the-loop* frameworks [307], offers a promising solution to enhance the robustness and clinical relevance of deep learning systems. Unlike traditional supervised learning, human-in-the-loop methods allow models to benefit from iterative feedback from human experts during both training and

deployment. For instance, in surgical robot assistance systems, human expert feedback enables models to learn recovery strategies for failure cases [308], allowing them to adapt and successfully complete tasks even after initial errors. However, balancing the effort required for human involvement with the need for model safety remains a significant challenge, particularly given the typically very long videos in clinical settings [2]. Future work should focus on developing efficient interfaces for expert-model interaction and exploring active learning strategies to prioritize cases where expert feedback is most impactful.

6.3.5 Developing Clinically Relevant Evaluation Metrics

Many current metrics for training and evaluating clinical video analysis models are adapted from general computer vision tasks, such as action recognition or segmentation, and may not fully align with the specific requirements of clinical applications [309]. These metrics often emphasize technical measures like pixel-level accuracy or area-under-curve scores, which, while important, may overlook clinically relevant factors such as procedural outcomes, decision-making efficacy, or the severity of errors. This misalignment underscores the need for metrics that better reflect the priorities and practical challenges of clinical settings. Clinically relevant metrics could incorporate subgroup analyses to differentiate between varying patient conditions, providing a more nuanced perspective on performance [310]. However, metrics developed within a single clinical site risk introducing personal and institutional biases. To address this limitation, future work could adopt consensus-driven approaches, such as the Delphi method, which gathers input from a diverse group of clinical experts to establish more generalizable and clinically meaningful evaluation frameworks [311].

6.3.6 Phased Real-world Verification

While the integration of clinical video analysis models with geometric feature estimation has shown promising results in controlled research environments from this thesis, deploying these models in real-world clinical scenarios requires rigorous validation. Real-world verification involves testing models across diverse settings to ensure they generalize effec-

tively beyond the training data [312]. This process should include large-scale, multi-center trials to evaluate performance across various institutions, patient demographics, and equipment setups [313]. However, such comprehensive evaluations may be challenging to implement initially due to resource constraints or logistical complexities. To address this, a phased verification approach could be encouraged, which involves progressively testing and refining models through increasingly complex and realistic scenarios, such as transitioning from simulation environments to human body samples [314], followed by small-scale clinical trials, and ultimately large-scale multi-center studies.

Bibliography

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. xiii, 4, 21, 22

[2] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2016. xiii, xv, 6, 8, 22, 25, 26, 29, 58, 63, 66, 72, 128

[3] Y. Kong and Y. Fu, "Close human interaction recognition using patch-aware models," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 167–178, 2015. xiii, 23, 24

[4] D. Rivoir, S. Bodenstedt, I. Funke, F. v. Bechtolsheim, M. Distler, J. Weitz, and S. Speidel, "Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 752–762, Springer, 2020. xiii, 6, 8, 23, 27, 32, 56, 57, 59, 70, 71, 72, 74, 75, 76, 77, 78

[5] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020. xiii, 24, 25, 44

[6] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific Data*, vol. 7, no. 1, p. 283, 2020. xiv, xvi, 27, 86, 92, 94, 95, 96

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015. xiv, 33, 47

[8] J. Huang, P. Zhu, M. Geng, J. Ran, X. Zhou, C. Xing, P. Wan, and X. Ji, "Range scaling global u-net for perceptual image enhancement on mobile devices," in *Proceedings of the European Conference on Computer Vision Workshops*, pp. 0–0, 2018. xiv, 33, 34

[9] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, vol. 51, pp. 2007–2028, 2020. xiv, 35

[10] S. Tukra, H. J. Marcus, and S. Giannarou, "See-through vision with unsupervised scene occlusion reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3779–3790, 2021. xiv, 6, 10, 27, 34, 35, 36, 37, 49, 85, 93

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. xv, 60

[12] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *Siam Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, 2014. xvi, 34, 92, 93, 95

[13] R. Daher, F. Vasconcelos, and D. Stoyanov, "A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence," *Medical Image Analysis*, p. 102994, 2023. xvi, 1, 10, 28, 32, 34, 36, 37, 44, 50, 82, 84, 85, 86, 87, 92, 93, 95, 99, 100

[14] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *Medical Image Analysis*, vol. 77, p. 102338, 2022. xvii, 10, 28, 45, 47, 48, 50, 84, 90, 92, 96, 97, 98, 100

[15] W. Xu, D. Xiang, G. Wang, R. Liao, M. Shao, and K. Li, "Multiview video-based 3-d pose estimation of patients in computer-assisted rehabilitation environment (caren)," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 196–206, 2022. xviii, 29, 45, 47, 110

[16] C. Loukas, "Video content analysis of surgical procedures," *Surgical endoscopy*, vol. 32, pp. 553–568, 2018. 1, 4

[17] K. Yao, N. Uedo, T. Kamada, T. Hirasawa, T. Nagahama, S. Yoshinaga, M. Oka, K. Inoue, K. Mabe, T. Yao, *et al.*, "Guidelines for endoscopic diagnosis of early gastric cancer," *Digestive Endoscopy*, vol. 32, no. 5, pp. 663–698, 2020. 1

[18] H. Forbes, F. I. Oprescu, T. Downer, N. M. Phillips, L. McTier, B. Lord, N. Barr, K. Alla, P. Bright, J. Dayton, *et al.*, "Use of videos to support teaching and learning of clinical skills in nursing education: A review," *Nurse Education Today*, vol. 42, pp. 53–56, 2016. 1

[19] D. Miskovic, M. Ni, S. M. Wyles, A. Parvaiz, and G. B. Hanna, "Observational clinical human reliability analysis (ochra) for competency assessment in laparoscopic colorectal surgery at the specialist level," *Surgical Endoscopy*, vol. 26, pp. 796–803, 2012. 1

[20] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–9, 2021. 1, 3

[21] K. Yuan, M. Holden, S. Gao, and W. Lee, "Anticipation for surgical workflow through instrument interaction and recognized signals," *Medical Image Analysis*, vol. 82, p. 102611, 2022. 1, 2, 5, 6, 8, 27, 29, 30, 31, 32, 51, 52, 55, 56, 57, 58, 59, 62, 64, 67, 70, 72, 73, 74, 75, 76, 78, 126, 127

[22] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 468–477, 2020. 1, 5, 53, 103, 109, 111, 113, 114, 115

[23] L. R. Kennedy-Metz, P. Mascagni, A. Torralba, R. D. Dias, P. Perona, J. A. Shah, N. Padoy, and M. A. Zenati, "Computer vision in the operating room: Opportunities and caveats," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 1, pp. 2–10, 2020. 1, 62

[24] J. Li, J. Xue, R. Cao, X. Du, S. Mo, K. Ran, and Z. Zhang, "Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3184–3193, 2024. 2, 4

[25] J. Liu, H. Wang, K. Stawarz, S. Li, Y. Fu, and H. Liu, "Vision-based human action quality assessment: A systematic review," *Expert Systems with Applications*, p. 125642, 2024. 2

[26] E. Elyan, P. Vuttipittayamongkol, P. Johnston, K. Martin, K. McPherson, C. F. Moreno-García, C. Jayne, and M. M. K. Sarker, "Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward," *Artificial Intelligence Surgery*, vol. 2, no. 1, pp. 24–45, 2022. 2, 22, 45

[27] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "De-smokegcn: generative cooperative networks for joint surgical smoke detection and removal," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1615–1625, 2019. 2, 4, 44

[28] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021. 2, 10, 32, 33, 34

[29] M. Javaid, A. Haleem, R. P. Singh, and M. Ahmed, "Computer vision to enhance healthcare domain: An overview of features, implementation, and opportunities," *Intelligent Pharmacy*, 2024. 2

[30] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2020. 2, 3

[31] G. Marcus, "The next decade in ai: four steps towards robust artificial intelligence," *arXiv preprint arXiv:2002.06177*, 2020. 2

[32] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017. 2

[33] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 283–291, 2018. 2

[34] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conference on Applications of Computer Vision*, pp. 148–157, IEEE, 2017. 2

[35] K. C. Demir, H. Schieber, T. Weise, D. Roth, M. May, A. Maier, and S. H. Yang, "Deep learning in surgical workflow analysis: a review of phase and step recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 11, pp. 5405–5417, 2023. 2, 7, 22, 23, 26, 29, 30

[36] Q. Abbas, M. E. Ibrahim, and M. A. Jaffar, "Video scene analysis: an overview and challenges on deep learning algorithms," *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 20415–20453, 2018. 3, 21

[37] M. Kolarik, M. Sarnovsky, J. Paralic, and F. Babic, "Explainability of deep learning models in medical video analysis: a survey," *PeerJ Computer Science*, vol. 9, p. e1253, 2023. 3

[38] W. Tang, P. M. van Ooijen, D. A. Sival, and N. M. Maurits, "Automatic two-dimensional & three-dimensional video analysis with deep learning for movement disorders: A systematic review," *Artificial Intelligence in Medicine*, p. 102952, 2024. 3

[39] I. Weinhold and S. Gurtner, "Understanding shortages of sufficient health care in rural areas," *Health Policy*, vol. 118, no. 2, pp. 201–214, 2014. 3

[40] M. C. Dohms, C. Collares, and I. C. Tibério, "Video-based feedback using real consultations for a formative assessment in communication skills," *BMC Medical Education*, vol. 20, pp. 1–9, 2020. 4

[41] J. Guo and B. Li, "The application of medical artificial intelligence technology in rural areas of developing countries," *Health Equity*, vol. 2, no. 1, pp. 174–181, 2018. 4

[42] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, 2022. 4

[43] V. Bellemo, G. Lim, T. H. Rim, G. S. Tan, C. Y. Cheung, S. Sadda, M.-g. He, A. Tufail, M. L. Lee, W. Hsu, *et al.*, "Artificial intelligence screening for diabetic retinopathy: the real-world emerging application," *Current Diabetes Reports*, vol. 19, pp. 1–12, 2019. 4

[44] D. A. Hashimoto, G. Rosman, D. Rus, and O. R. Meireles, "Artificial intelligence in surgery: promises and perils," *Annals of Surgery*, vol. 268, no. 1, pp. 70–76, 2018. 4

[45] F. Nagi, R. Salih, M. Alzubaidi, H. Shah, T. Alam, Z. Shah, and M. Househ, "Applications of artificial intelligence (ai) in medical education: a scoping review," *Healthcare Transformation with Informatics and Artificial Intelligence*, pp. 648–651, 2023. 4

[46] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, and J. Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *Medical Image Analysis*, vol. 68, p. 101900, 2021. 4, 5, 36, 80, 85, 93, 100

[47] L. Li, R. A. Paris, C. Pinson, Y. Wang, J. Coco, J. Heard, J. A. Adams, D. V. Fabbri, and B. Bodenheimer, "Emergency clinical procedure detection with deep learning," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pp. 158–163, IEEE, 2020. 4

[48] H. Ali, M. Sharif, M. Yasmin, M. H. Rehmani, and F. Riaz, "A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract," *Artificial Intelligence Review*, vol. 53, pp. 2635–2707, 2020. 4, 5, 41, 42

[49] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, pp. 1–74, 2021. 5

[50] S. Xiong, Y. Tan, G. Wang, P. Yan, and X. Xiang, "Learning feature relationships in cnn model via relational embedding convolution layer," *Neural Networks*, vol. 179, p. 106510, 2024. 5

[51] H. Choi, H. Lee, S. Jeong, and D. Min, "Environment agnostic representation for visual reinforcement learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 263–273, 2023. 5

[52] W. Chen, Y. Liu, J. Hu, and Y. Yuan, "Dynamic depth-aware network for endoscopy super-resolution," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 5189–5200, 2022. 5, 52, 84, 90

[53] F. Sardari, B. Ommer, and M. Mirmehdi, "Unsupervised view-invariant human posture representation," in *Proceedings of the 32nd British Machine Vision Conference*, BMVA Press, 2021. 5

[54] A. Murali, D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, D. Mutter, and N. Padoy, "Latent graph representations for critical view of safety assessment," *IEEE Transactions on Medical Imaging*, 2023. 5, 51

[55] T. Powers, E. Hatamimajoumerd, W. Chu, V. Rajendran, R. Shah, F. Diabour, M. Vaillant, R. Fletcher, and S. Ostadabbas, "Vision-based treatment localization with limited data: Automated documentation of military emergency medical procedures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1819–1828, 2023. 5

[56] L. Shen, W. Zhao, D. Capaldi, J. Pauly, and L. Xing, "A geometry-informed deep learning framework for ultra-sparse 3d tomographic image reconstruction," *Computers in Biology and Medicine*, vol. 148, p. 105710, 2022. 5

[57] D.-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. C. Niebles, "What makes a video a video: Analyzing temporal information in video understanding models and datasets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7366–7375, 2018. 5

[58] C.-Y. Wu and P. Krahenbuhl, "Towards long-form video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1884–1894, 2021. 5

[59] X. Liu, M. Stiber, J. Huang, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pp. 3–13, Springer, 2020. 6

[60] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17562–17571, 2022. 6, 36

[61] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial-temporal transformations for video inpainting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 528–543, Springer, 2020. 6, 34, 35, 36, 37, 86, 88, 89, 92, 93

[62] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2020. 6, 7

[63] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021. 7, 26

[64] A. J. Hung, R. Bao, I. O. Sunmola, D.-A. Huang, J. H. Nguyen, and A. Anandkumar, "Capturing fine-grained details for video-based automation of suturing skills assessment," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 3, pp. 545–552, 2023. 7

[65] Z. Cui, R. Ma, C. H. Yang, A. Malpani, T. N. Chu, A. Ghazi, J. W. Davis, B. J. Miles, C. Lau, Y. Liu, *et al.*, "Capturing relationships between suturing sub-skills to improve automatic suturing assessment," *NPJ Digital Medicine*, vol. 7, no. 1, p. 152, 2024. 7

[66] B. Debnath, M. O'brien, M. Yamaguchi, and A. Behera, "A review of computer vision-based approaches for physical rehabilitation and assessment," *Multimedia Systems*, vol. 28, no. 1, pp. 209–239, 2022. 7, 45

[67] T. Karácsony, L. A. Jeni, F. De la Torre, and J. P. S. Cunha, "Deep learning methods for single camera based clinical in-bed movement action recognition," *Image and Vision Computing*, vol. 143, p. 104928, 2024. 7

[68] S. Wang, D. Yang, P. Zhai, Q. Yu, T. Suo, Z. Sun, K. Li, and L. Zhang, "A survey of video-based action quality assessment," in *2021 International Conference on Networking Systems of AI*, pp. 1–9, IEEE, 2021. 7, 11, 37, 38, 102, 113

[69] A. E. Abdelaal, A. Avinash, M. Kalia, G. D. Hager, and S. E. Salcudean, "A multi-camera, multi-view system for training and skill assessment for robot-assisted surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 8, pp. 1369–1377, 2020. 7, 39

[70] P. Mascagni, D. Alapatt, L. Sestini, M. S. Altieri, A. Madani, Y. Watanabe, A. Alseidi, J. A. Redan, S. Alfieri, G. Costamagna, *et al.*, "Computer vision in surgery: from potential to clinical value," *NPJ Digital Medicine*, vol. 5, no. 1, p. 163, 2022. 7

[71] N. J. Cronin, "Using deep neural networks for kinematic analysis: Challenges and opportunities," *Journal of Biomechanics*, vol. 123, p. 110460, 2021. 7

[72] S. Protserov, J. Hunter, H. Zhang, P. Mashouri, C. Masino, M. Brudno, and A. Madani, "Development, deployment and scaling of operating room-ready artificial intelligence for real-time surgical decision support," *NPJ Digital Medicine*, vol. 7, no. 1, p. 231, 2024. 8

[73] K. Yuan, M. Holden, S. Gao, and W.-S. Lee, "Surgical workflow anticipation using instrument interaction," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 615–625, Springer, 2021. 9, 32, 57, 72

[74] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, *et al.*, "Surgical data science for next-generation interventions," *Nature Biomedical Engineering*, vol. 1, no. 9, pp. 691–696, 2017. 9, 56

[75] Y. Ban, G. Rosman, T. Ward, D. Hashimoto, T. Kondo, H. Iwaki, O. Meireles, and D. Rus, "Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows," in *2021 IEEE International Conference on Robotics and Automation*, pp. 14531–14538, IEEE, 2021. 9, 24

[76] J. D. Mason, J. Ansell, N. Warren, and J. Torkington, "Is motion analysis a valid tool for assessing laparoscopic skill?," *Surgical Endoscopy*, vol. 27, pp. 1468–1477, 2013. 11

[77] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 11

[78] K. E. Weiss, M. Kolbe, A. Nef, B. Grande, B. Kalirajan, M. Meboldt, and Q. Lohmeyer, "Data-driven resuscitation training using pose estimation," *Advances in Simulation*, vol. 8, no. 1, p. 12, 2023. 11

[79] T. Igaki, D. Kitaguchi, H. Matsuzaki, K. Nakajima, S. Kojima, H. Hasegawa, N. Takeshita, Y. Kinugasa, and M. Ito, "Automatic surgical skill assessment system based on concordance of standardized surgical field development using artificial intelligence," *JAMA Surgery*, vol. 158, no. 8, pp. e231131–e231131, 2023. 11

[80] K. Lam, J. Chen, Z. Wang, F. M. Iqbal, A. Darzi, B. Lo, S. Purkayastha, and J. M. Kinross, "Machine learning for technical skill assessment in surgery: a systematic review," *NPJ Digital Medicine*, vol. 5, no. 1, p. 24, 2022. 11, 39

[81] Z. Zhong, M. Martin, M. Voit, J. Gall, and J. Beyerer, "A survey on deep learning techniques for action anticipation," *arXiv preprint arXiv:2309.17257*, 2023. 20, 31

[82] A. Huaulmé, F. Despinoy, S. A. H. Perez, K. Harada, M. Mitsuishi, and P. Jannin, "Automatic annotation of surgical activities using virtual reality environments," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, pp. 1663–1671, 2019. 20

[83] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1491–1498, IEEE, 2009. 20

[84] A. Abelsson, C. Gwinnutt, P. Greig, J. Smart, and K. Mackie, "Validating peer-led assessments of cpr performance," *Resuscitation Plus*, vol. 3, p. 100022, 2020. 20

[85] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7099–7122, 2022. 20

[86] A. Anwar, S. Kanwal, M. Tahir, M. Saqib, M. Uzair, M. K. I. Rahmani, and H. Ullah, "Image aesthetic assessment: a comparative study of hand-crafted & deep learning models," *IEEE Access*, vol. 10, pp. 101770–101789, 2022. 20

[87] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288–303, 2008. 20

[88] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: A survey," *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, 2015. 21

[89] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019. 21

[90] F. Zhang, W. Li, Y. Zhang, and Z. Feng, "Data driven feature selection for machine learning algorithms in computer vision," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4262–4272, 2018. 21

[91] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 21

[92] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. 21

[93] D. Ouyang, Z. Wu, B. He, and J. Zou, "Deep learning for biomedical videos: perspective and recommendations," in *Artificial Intelligence in Medicine*, pp. 37–48, Elsevier, 2021. 22

[94] M. Tabish, Z.-u.-R. Tanooli, and M. Shaheen, "Activity recognition framework in sports videos," *Multimedia Tools and Applications*, pp. 1–23, 2024. 22

[95] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognition Letters*, vol. 118, pp. 14–22, 2019. 22

[96] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Video processing using deep learning techniques: A systematic literature review," *IEEE Access*, vol. 9, pp. 139489–139507, 2021. 23

[97] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 23, 53, 113

[98] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017. 24

[99] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. 24

[100] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12922–12943, 2023. 24

[101] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023. 24

[102] Y. Jing and F. Wang, "Tp-vit: A two-pathway vision transformer for video action recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2185–2189, IEEE, 2022. 24

[103] F. D. Keles, P. M. Wijewardena, and C. Hegde, "On the computational complexity of self-attention," in *International Conference on Algorithmic Learning Theory*, pp. 597–619, PMLR, 2023. 25, 30

[104] Z. Luo, Y. Xiao, F. Yang, J. T. Zhou, and Z. Fang, "Rhythmer: Ranking-based skill assessment with rhythm-aware transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 25

[105] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Medical Image Analysis*, vol. 85, p. 102762, 2023. 25

[106] T. J. Saun, K. J. Zuo, and T. P. Grantcharov, "Video technologies for recording open surgery: a systematic review," *Surgical Innovation*, vol. 26, no. 5, pp. 599–612, 2019. 26

[107] O. Özgüner, T. Shkurti, S. Huang, R. Hao, R. C. Jackson, W. S. Newman, and M. C. Çavuşoğlu, "Camera-robot calibration for the da vinci robotic surgery system," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 4, pp. 2154–2161, 2020. 27

[108] Z. Fu, Z. Jin, C. Zhang, Z. He, Z. Zha, C. Hu, T. Gan, Q. Yan, P. Wang, and X. Ye, "The future of endoscopic navigation: A review of advanced endoscopic vision technology," *IEEE Access*, vol. 9, pp. 41144–41167, 2021. 27, 48, 52, 84, 86, 88

[109] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng, "Temporal memory relation network for workflow recognition from surgical video," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1911–1923, 2021. 27

[110] H. Li, X. Hou, R. Lin, M. Fan, S. Pang, L. Jiang, Q. Liu, and L. Fu, "Advanced endoscopic methods in gastrointestinal diseases: a systematic review," *Quantitative Imaging in Medicine and Surgery*, vol. 9, no. 5, p. 905, 2019. 28

[111] H. Hautmann, J. Hetzel, R. Eberhardt, F. Stanzel, M. Wagner, A. Schneider, R. Dirschinger, and A. Poszler, "Cross-sectional survey on bronchoscopy in germany–the current status of clinical practice," *Pneumologie*, vol. 70, no. 02, pp. 110–116, 2016. 28

[112] P. Valdastri, M. Simi, and R. J. Webster III, "Advanced technologies for gastrointestinal endoscopy," *Annual Review of Biomedical Engineering*, vol. 14, no. 1, pp. 397–429, 2012. 28

[113] A. A. Mathews, P. V. Draganov, and D. Yang, "Endoscopic management of colorectal polyps: From benign to malignant polyps," *World Journal of Gastrointestinal Endoscopy*, vol. 13, no. 9, p. 356, 2021. 28

[114] A. L. Faulx, S. Kothari, R. D. Acosta, D. Agrawal, D. H. Bruining, V. Chandrasekhara, M. A. Eloubeidi, R. D. Fanelli, S. R. Gurudu, M. A. Khashab, *et al.*, "The role of endoscopy in subepithelial lesions of the gi tract," *Gastrointestinal Endoscopy*, vol. 85, no. 6, pp. 1117–1132, 2017. 28

[115] B. Chatrangsun and R.-K. Vilaichone, "Endoscopic diagnosis for h. pylori infection: white light imaging (wli) vs. image-enhanced endoscopy (iee)," *Asian Pacific Journal of Cancer Prevention: APJCP*, vol. 22, no. 9, p. 3031, 2021. 28

[116] S. Sunny, G. Cheng, D. Daniel, P. Lo, S. Ochoa, C. Howell, N. Vogel, A. Majid, and J. Aizenberg, "Transparent antifouling material for improved operative field visibility in endoscopy," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. 11676–11681, 2016. 28

[117] M. T. Parks, Z. Wang, and K.-C. Siu, "Current low-cost video-based motion analysis options for clinical rehabilitation: a systematic review," *Physical Therapy*, vol. 99, no. 10, pp. 1405–1425, 2019. 28

[118] R. Stone, M. Cooke, and M. Mitchell, "Undergraduate nursing students' use of video technology in developing confidence in clinical skills for practice: A systematic integrative literature review," *Nurse Education Today*, vol. 84, p. 104230, 2020. 28

[119] H. Gattinger, M. Stolt, V. Hantikainen, S. Koepke, B. Senn, and H. Leino-Kilpi, "A systematic review of observational instruments used to assess nurses' skills in patient mobilisation," *Journal of Clinical Nursing*, vol. 24, no. 5-6, pp. 640–661, 2015. 29

[120] S. Heinerichs, N. M. Cattano, and K. E. Morrison, "Assessing nonverbal communication skills through video recording and debriefing of clinical skill simulation exams," *Athletic Training Education Journal*, vol. 8, no. 1-2, pp. 17–22, 2013. 29

[121] P. Yeates, A. Moult, J. Lefroy, J. Walsh-House, L. Clews, R. McKinley, and R. Fuller, "Understanding and developing procedures for video-based assessment in medical education," *Medical Teacher*, vol. 42, no. 11, pp. 1250–1260, 2020. 29

[122] C. F. Mackenzie and Y. Xiao, "Video techniques and data compared with observation in emergency trauma care," *BMJ Quality & Safety*, vol. 12, no. suppl 2, pp. ii51–ii57, 2003. 29

[123] X. Hu, J. Dai, M. Li, C. Peng, Y. Li, and S. Du, "Online human action detection and anticipation in videos: A survey," *Neurocomputing*, vol. 491, pp. 395–413, 2022. 29

[124] A. Garcia-Martinez, J. M. Vicente-Samper, and J. M. Sabater-Navarro, "Automatic detection of surgical haemorrhage using computer vision," *Artificial Intelligence in Medicine*, vol. 78, pp. 55–60, 2017. 29

[125] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8739–8748, 2018. 30

[126] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "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*, pp. 47–54, Springer, 2016. 30

[127] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584, 2019. 30

[128] F. Sener, D. Singhania, and A. Yao, "Temporal aggregate representations for long-range video understanding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 154–171, Springer, 2020. 31, 74, 75, 76, 77, 78

[129] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 729–738, 2013. 31

[130] C. Shi, Y. Zheng, and A. M. 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*, pp. 8017–8024, IEEE, 2022. 31

[131] H. Kossowsky and I. Nisky, "Predicting the timing of camera movements from the kinematics of instruments in robotic-assisted surgery using artificial neural networks," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 2, pp. 391–402, 2022. 31

[132] J. Zhang, S. Zhou, Y. Wang, S. Shi, C. Wan, H. Zhao, X. Cai, and H. Ding, "Laparoscopic image-based critical action recognition and anticipation with explainable features," *IEEE Journal of Biomedical and Health Informatics*, 2023. 31, 46, 52, 126

[133] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1069–1078, 2018. 32, 56, 58, 74, 75, 76, 77, 78

[134] A. Marafioti, M. Hayoz, M. Gallardo, P. Márquez Neila, S. Wolf, M. Zinkernagel, and R. Sznitman, "Catanet: Predicting remaining cataract surgery duration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 426–435, Springer, 2021. 32, 56, 73, 74, 75, 77, 79

[135] X. Zhang, N. Al Moubayed, and H. P. Shum, "Towards graph representation learning based surgical workflow anticipation," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics*, pp. 01–04, IEEE, 2022. 32, 57, 58, 59, 62, 64, 65, 68, 70, 74, 75, 76, 77, 78

[136] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 586–593, Springer, 2017. 32, 56, 74, 75, 76, 77, 78

[137] Z. Yang, J. Dai, and J. Pan, "3d reconstruction from endoscopy images: A survey," *Computers in Biology and Medicine*, p. 108546, 2024. 33, 48

[138] R. Al Sobbahi and J. Tekli, "Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: Overview, empirical evaluation, and challenges," *Signal Processing: Image Communication*, vol. 109, p. 116848, 2022. 34

[139] C. Zhang, N. Zhang, D. Wang, Y. Cao, and B. Liu, "Artifact detection in endoscopic video with deep convolutional neural networks," in *2020 Second International Conference on Transdisciplinary AI*, pp. 1–8, IEEE, 2020. 34

[140] W. Quan, J. Chen, Y. Liu, D.-M. Yan, and P. Wonka, "Deep learning-based image and video inpainting: A survey," *International Journal of Computer Vision*, vol. 132, no. 7, pp. 2367–2400, 2024. 34, 35, 84

[141] M. Arnold, A. Ghosh, S. Ameling, and G. Lacey, "Automatic segmentation and inpainting of specular highlights for endoscopic imaging," *EURASIP Journal on Image and Video Processing*, vol. 2010, pp. 1–12, 2010. 34, 92, 93

[142] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014. 35

[143] S. Zhou, C. Li, K. C. Chan, and C. C. Loy, "Propainter: Improving propagation and transformer for video inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10477–10486, 2023. 36, 91

[144] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5792–5801, 2019. 36

[145] D.-F. Shen, J.-J. Guo, G.-S. Lin, and J.-Y. Lin, "Content-aware specular reflection suppression based on adaptive image inpainting and neural network for endoscopic images," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105414, 2020. 36, 85

[146] D. He, Y. Li, L. Chen, X. Xiao, Y. Xue, Z. Wang, and Y. Li, "Dual-guided network for endoscopic image segmentation with region and boundary cues," *Biomedical Signal Processing and Control*, vol. 91, p. 106059, 2024. 37

[147] N. Louis, L. Zhou, S. J. Yule, R. D. Dias, M. Manojlovich, F. D. Pagani, D. S. Likosky, and J. J. Corso, "Temporally guided articulated hand pose tracking in surgical videos," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 1, pp. 117–125, 2023. 38, 39

[148] C.-H. Chi, J.-Y. Tsou, and F.-C. Su, "Effects of rescuer position on the kinematics of cardiopulmonary resuscitation (cpr) and the force of delivered compressions," *Resuscitation*, vol. 76, no. 1, pp. 69–75, 2008. 38

[149] L. Xu, H. Gong, Y. Zhong, F. Wang, S. Wang, L. Lu, J. Ding, C. Zhao, W. Tang, and J. Xu, "Real-time monitoring of manual acupuncture stimulation parameters based on domain adaptive 3d hand pose estimation," *Biomedical Signal Processing and Control*, vol. 83, p. 104681, 2023. 38

[150] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. Peng, "Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14628–14637, 2024. 38, 53

[151] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6331–6340, 2019. 38

[152] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9839–9848, 2020. 38

[153] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6057–6066, 2018. 38

[154] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7862–7871, 2019. 38

[155] S. Sardari, S. Sharifzadeh, A. Daneshkhah, B. Nakisa, S. W. Loke, V. Palade, and M. J. Duncan, "Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review," *Computers in Biology and Medicine*, vol. 158, p. 106835, 2023. 39, 48, 107

[156] Y. Liu, Z. Zhao, P. Shi, and F. Li, "Towards surgical tools detection and operative skill assessment based on deep learning," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 1, pp. 62–71, 2022. 39, 42, 45

[157] Z. Li, L. Gu, W. Wang, R. Nakamura, and Y. Sato, "Surgical skill assessment via video semantic aggregation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 410–420, Springer, 2022. 39

[158] S. Deb, M. F. Islam, S. Rahman, and S. Rahman, "Graph convolutional networks for assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 410–419, 2022. 39, 53, 54, 103, 113, 114

[159] L. Yao, Q. Lei, H. Zhang, J. Du, and S. Gao, "A contrastive learning network for performance metric and assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 39, 103, 109, 111

[160] H. Wang, K. Sugand, S. Newman, G. Jones, J. Cobb, and E. Auvinet, "Are multiple views superior to a single view when teaching hip surgery? a single-blinded randomized controlled trial of technical skill acquisition," *PloS One*, vol. 14, no. 1, p. e0209904, 2019. 39

[161] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, *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, 2014. 40

[162] K. Schoeffmann, M. Taschwer, S. Sarny, B. Münzer, M. J. Primus, and D. Putzgruber, "Cataract-101: video dataset of 101 cataract surgeries," in *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 421–425, ACM, 2018. 40, 58, 63, 73

[163] A. Zia, K. Bhattacharyya, X. Liu, Z. Wang, M. Berniker, S. Kondo, E. Colleoni, D. Psychogyios, Y. Jin, J. Zhou, *et al.*, "Objective surgical skills assessment and tool localization: Results from the miccai 2021 simsurgskill challenge," *arXiv preprint arXiv:2212.04448*, 2022. 40

[164] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek, *et al.*, "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark," *Medical Image Analysis*, vol. 86, p. 102770, 2023. 40

[165] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, p. 2, 2018. 40

[166] A. Das, B. Sidiqi, L. Mennillo, Z. Mao, M. Brudfors, M. Xochicale, D. Z. Khan, N. Newall, J. G. Hanrahan, M. J. Clarkson, *et al.*, "Automated surgical skill assessment in endoscopic pituitary surgery using real-time instrument tracking on a high-fidelity bench-top phantom," *Healthcare Technology Letters*, vol. 11, no. 6, pp. 336–344, 2024. 40

[167] M. Devanne, O. R. Neris, M. Lempereur, A. Thepaut, *et al.*, "A medical low-back pain physical rehabilitation database for human body movement analysis," in *2024 International Joint Conference on Neural Networks*, pp. 1–8, IEEE, 2024. 40

[168] S. Wang, S. Wang, D. Yang, M. Li, H. Kuang, X. Zhao, L. Su, P. Zhai, and L. Zhang, "Cpr-coach: Recognizing composite error actions based on single-class training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18782–18792, 2024. 39

[169] T. Georgiou, Y. Liu, W. Chen, and M. Lew, "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *International Journal of Multimedia Information Retrieval*, vol. 9, pp. 135–170, 2020. 41

[170] M. Abdullahi, O. N. Oyelade, A. F. D. Kana, M. A. Bagiwa, F. B. Abdullahi, S. B. Junaidu, I. Iliyasu, A. Ore-ofe, and H. Chiroma, "A systematic literature review of visual feature learning: deep learning techniques, applications, challenges and future directions," *Multimedia Tools and Applications*, pp. 1–58, 2024. 41

[171] T. Kalinke, C. Tzomakas, and W. von Seelen, "A texture-based object detection and an adaptive model-based classification," in *Procs. IEEE Intelligent Vehicles Symposium*, vol. 98, pp. 341–346, Citeseer, 1998. 42

[172] J. Yuan, D. Wang, and R. Li, "Remote sensing image segmentation by combining spectral and texture features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 16–24, 2013. 42

[173] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. Lau, and C. C. Poon, "Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 41–47, 2016. 42

[174] V. Shereena, J. M. David, *et al.*, "Content based image retrieval: A review," in *Computer Science & Information Technology, Computer Science Conference Proceedings (CSCP)*, pp. 65–77, 2014. 42

[175] A. Humeau-Heurtier, "Texture feature extraction methods: A survey," *IEEE Access*, vol. 7, pp. 8975–9000, 2019. 42

[176] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, Ieee, 1999. 42

[177] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pp. 404–417, Springer, 2006. 43

[178] A. Humeau-Heurtier, "Color texture analysis: A survey," *IEEE Access*, vol. 10, pp. 107993–108003, 2022. 43

[179] L. Liu, P. Fieguth, X. Wang, M. Pietikäinen, and D. Hu, "Evaluation of lbp and deep texture descriptors with a new robustness benchmark," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 69–86, Springer, 2016. 43

[180] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, Springer, 2014. 43

[181] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. 43, 107

[182] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, Computational and Biological Learning Society, 2015. 43

[183] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015. 43

[184] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017. 43

[185] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, pp. 10347–10357, PMLR, 2021. 44

[186] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. 44

[187] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009. 44

[188] N. Xi, J. Meng, and J. Yuan, "Forest graph convolutional network for surgical action triplet recognition in endoscopic videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8550–8561, 2022. 44

[189] X. Pennec, "Toward a generic framework for recognition based on uncertain geometric features," *Videre: Journal of Computer Vision Research*, vol. 1, no. 2, pp. 58–87, 1998. 44

[190] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Tenth IEEE International Conference on Computer Vision*, vol. 1, pp. 654–661, IEEE, 2005. 44

[191] V. Kanhangad, A. Kumar, and D. Zhang, "Combining 2d and 3d hand geometry features for biometric verification," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 39–44, IEEE, 2009. 45

[192] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *2018 IEEE Winter Conference on Applications of Computer Vision*, pp. 691–699, IEEE, 2018. 45, 46, 63

[193] I. Laina, N. Rieke, C. Rupprecht, J. P. Vizcaíno, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," in *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pp. 664–672, Springer, 2017. 45

[194] L. Xi, Y. Zhao, L. Chen, Q. H. Gao, W. Tang, T. R. Wan, and T. Xue, "Recovering dense 3d point clouds from single endoscopic image," *Computer Methods and Programs in Biomedicine*, vol. 205, p. 106077, 2021. 45

[195] J. Wang, Y. Jin, S. Cai, H. Xu, P.-A. Heng, J. Qin, and L. Wang, "Real-time landmark detection for precise endoscopic submucosal dissection via shape-aware relation network," *Medical Image Analysis*, vol. 75, p. 102291, 2022. 45

[196] T. Qiao, Q. Men, F. W. Li, Y. Kubotani, S. Morishima, and H. P. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *European Conference on Computer Vision*, pp. 474–491, Springer, 2022. 45, 82

[197] A. Bas and W. A. Smith, "What does 2d geometric information really tell us about 3d face shape?," *International Journal of Computer Vision*, vol. 127, pp. 1455–1473, 2019. 45

[198] I. Barua, D. G. Vinsard, H. C. Jodal, M. Løberg, M. Kalager, Ø. Holme, M. Misawa, M. Bretthauer, and Y. Mori, "Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis," *Endoscopy*, vol. 53, no. 03, pp. 277–284, 2021. 45

[199] Y. Wang, Q. Sun, Z. Liu, and L. Gu, "Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art," *Robotics and Autonomous Systems*, vol. 149, p. 103945, 2022. 45, 57

[200] A. Madani, B. Namazi, M. S. Altieri, D. A. Hashimoto, A. M. Rivera, P. H. Pucher, A. Navarrete-Welton, G. Sankaranarayanan, L. M. Brunt, A. Okrainec, *et al.*, "Artificial

intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy," *Annals of surgery*, vol. 276, no. 2, pp. 363–369, 2022. 45

[201] K. Lam, F. P.-W. Lo, Y. An, A. Darzi, J. M. Kinross, S. Purkayastha, and B. Lo, "Deep learning for instrument detection and assessment of operative skill in surgical videos," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 4, pp. 1068–1071, 2022. 45

[202] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, 2018. 45

[203] A. Fukuta, S. Yamashita, J. Maniwa, A. Tamaki, T. Kondo, N. Kawakubo, K. Nagata, T. Matsuura, and T. Tajiri, "Artificial intelligence facilitates the potential of simulator training: An innovative laparoscopic surgical skill validation system using artificial intelligence technology," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–7, 2024. 45, 46

[204] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017. 45

[205] D. Sarikaya and P. Jannin, "Towards generalizable surgical activity recognition using spatial temporal graph convolutional networks," *arXiv preprint arXiv:2001.03728*, 2020. 46, 58

[206] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "Deeplabcut: markerless pose estimation of user-defined body parts with deep learning," *Nature Neuroscience*, vol. 21, no. 9, pp. 1281–1289, 2018. 46

[207] E. Bermejo, K. Taniguchi, Y. Ogawa, R. Martos, A. Valsecchi, P. Mesejo, O. Ibáñez, and K. Imaizumi, "Automatic landmark annotation in 3d surface scans of skulls: Methodological proposal and reliability study," *Computer Methods and Programs in Biomedicine*, vol. 210, p. 106380, 2021. 46

[208] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023. 46, 50

[209] P. Zeng, S. Liu, S. He, Q. Zheng, J. Wu, Y. Liu, G. Lyu, and P. Liu, "Tuspm-net: A multi-task model for thyroid ultrasound standard plane recognition and detection of key anatomical structures of the thyroid," *Computers in Biology and Medicine*, vol. 163, p. 107069, 2023. 46

[210] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016. 46, 64

[211] B. Pande, K. Padamwar, S. Bhattacharya, S. Roshan, and M. Bhamare, "A review of image annotation tools for object detection," in *2022 International Conference on Applied Artificial Intelligence and Computing*, pp. 976–982, IEEE, 2022. 46

[212] M. G. Ragab, S. J. Abdulkadir, A. Muneer, A. Alqushaibi, E. H. Sumiea, R. Qureshi, S. M. Al-Selwi, and H. Alhussian, "A comprehensive systematic review of yolo for medical object detection (2018 to 2023)," *IEEE Access*, vol. 12, pp. 57815–57836, 2024. 46

[213] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021. 46

[214] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022. 47

[215] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020. 47

[216] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 4, pp. 1–47, 2020. 47

[217] C. Fifelski, A. Brinkmann, S. M. Ortmann, M. Isken, and A. Hein, "Multi depth camera system for 3d data recording for training and education of nurses," in *2018 International Conference on Computational Science and Computational Intelligence*, pp. 679–684, IEEE, 2018. 47

[218] A. K. Ingale *et al.*, "Real-time 3d reconstruction techniques applied in dynamic scenes: A systematic literature review," *Computer Science Review*, vol. 39, p. 100338, 2021. 47

[219] C. Brambilla, R. Marani, L. Romeo, M. L. Nicora, F. A. Storm, G. Reni, M. Malosio, T. D'Orazio, and A. Scano, "Azure kinect performance evaluation for human motion and upper limb biomechanical analysis," *Heliyon*, vol. 9, no. 11, 2023. 47, 51

[220] A. Vakunov, C.-L. Chang, F. Zhang, G. Sung, M. Grundmann, and V. Bazarevsky, "Mediapipe hands: On-device real-time hand tracking," in *Workshop on Computer Vision for AR/VR*, vol. 2, p. 5, 2020. 47, 107

[221] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, 2021. 48

[222] Z. Huang, Y. Wen, Z. Wang, J. Ren, and K. Jia, "Surface reconstruction from point clouds: A survey and a benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 48

[223] A. Shalaby, M. Elmogy, and A. A. El-Fetouh, "Algorithms and applications of structure from motion (sfm): A survey," *Algorithms*, vol. 6, no. 06, 2017. 48

[224] M. Hayoz, C. Hahne, M. Gallardo, D. Candinas, T. Kurmann, M. Allan, and R. Sznitman, "Learning how to robustly estimate camera pose in endoscopic videos," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 7, pp. 1185–1192, 2023. 49

[225] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. 49, 50

[226] K. T. Nguyen, F. Tozzi, N. Rashidian, W. Willaert, J. Vankerschaver, and W. De Neve, "Towards abdominal 3-d scene rendering from laparoscopy surgical videos using nerfs," in *International Workshop on Machine Learning in Medical Imaging*, pp. 83–93, Springer, 2023. 49

[227] B. G. Gerats, J. M. Wolterink, and I. A. Broeders, "Nerf-or: neural radiance fields for operating room scene reconstruction from sparse-view rgb-d videos," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10, 2024. 49

[228] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, *et al.*, "Neural 3d video synthesis from multi-view video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5521–5531, 2022. 49

[229] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023. 49, 50

[230] A. Furnari and G. M. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4021–4036, 2020. 52

[231] M. Liao, F. Lu, D. Zhou, S. Zhang, W. Li, and R. Yang, "Dvi: Depth guided video inpainting for autonomous driving," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pp. 1–17, Springer, 2020. 52, 86

[232] Y. Yamashita, K. Shimosato, and N. Ukita, "Boundary-aware image inpainting with multiple auxiliary cues," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 619–629, 2022. 52, 86, 91

[233] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017. 53, 69

[234] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018. 53, 113, 114

[235] K. Zheng, J. Wu, J. Zhang, and C. Guo, "A skeleton-based rehabilitation exercise assessment system with rotation invariance," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 53, 54, 109, 111, 113, 114, 115, 119, 120

[236] X. Wang and Z. Zhu, "Context understanding in computer vision: A survey," *Computer Vision and Image Understanding*, vol. 229, p. 103646, 2023. 57

[237] L. Li, X. Li, B. Ouyang, S. Ding, S. Yang, and Y. Qu, "Autonomous multiple instruments tracking for robot-assisted laparoscopic surgery with visual tracking space vector method," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 2, pp. 733–743, 2021. 57

[238] D. Rivoir, S. Bodenstedt, F. v. Bechtolsheim, M. Distler, J. Weitz, and S. Speidel, "Unsupervised temporal video segmentation as an auxiliary task for predicting the remaining surgery duration," in *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, pp. 29–37, Springer, 2019. 57

[239] Y. Wu, L. Zhu, X. Wang, Y. Yang, and F. Wu, "Learning to anticipate egocentric actions by imagination," *IEEE Transactions on Image Processing*, vol. 30, pp. 1143–1152, 2020. 57

[240] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, Z. Yifu, C. Wong, D. Montes, *et al.*, "ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation," *Zenodo*, 2022. 60, 63

[241] M. Fox, M. Taschwer, and K. Schoeffmann, "Pixel-based tool segmentation in cataract surgery videos with mask R-CNN," in *33rd IEEE International Symposium on Computer-Based Medical Systems*, pp. 565–568, IEEE, 2020. 63

[242] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022. 64, 69

[243] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 627–636, Springer, 2020. 65

[244] G. Mena, J. Snoek, S. Linderman, and D. Belanger, "Learning latent permutations with gumbel-sinkhorn networks," in *ICLR 2018 Conference Track*, 2018. 67

[245] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. 69

[246] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018. 71, 72

[247] L. Biewald, "Experiment tracking with weights and biases," 2020. Software available from wandb.com. 74

[248] J. Wu, R. Tao, and G. Zheng, "Nonlinear regression of remaining surgical duration via bayesian lstm-based deep negative correlation learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 421–430, Springer, 2022. 75

[249] B. Wang, L. Li, Y. Nakashima, R. Kawasaki, and H. Nagahara, "Real-time estimation of the remaining surgery duration for cataract surgery using deep convolutional neural networks and long short-term memory," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 80, 2023. 75

[250] M. S. Copenhaver, T. H. Friend, C. Fitzgerald-Brown, M. Fernandez, M. Addesa, J. Cassidy, M. Rosa, J. Ouellette, J. Plunkett, D. Spracklin, *et al.*, "Improving operating room and surgical instrumentation efficiency, safety, and communication via the implementation of emergency laparoscopic cholecystectomy and appendectomy conversion case carts," *Perioperative Care and Operating Room Management*, vol. 8, pp. 33–37, 2017. 76

[251] E. Checcucci, S. De Cillis, D. Amparore, V. Gabriele, F. Piramide, A. Piana, C. Fiori, P. Piazzolla, and F. Porpiglia, "Artificial intelligence alert systems during robotic surgery: A new potential tool to improve the safety of the intervention," *Urology Video Journal*, vol. 18, p. 100221, 2023. 76

[252] E. Shaw and B. C. Patel, "Complicated cataract," in *StatPearls [Internet]*, StatPearls Publishing, 2023. 79

[253] Z. Chang, G. A. Koulieris, and H. P. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv preprint arXiv:2306.04542*, 2023. 82

[254] R. Yang and Y. Yu, "Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis," *Frontiers in Oncology*, vol. 11, p. 638182, 2021. 84

[255] Y. Ge, Y. Xiao, Z. Xu, X. Wang, and L. Itti, "Contributions of shape, texture, and color in visual recognition," in *European Conference on Computer Vision*, pp. 369–386, Springer, 2022. 84

[256] D. Schnieders and K.-Y. K. Wong, "Camera and light calibration from reflections on a sphere," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1536–1547, 2013. 84

[257] Y. Wang, H. Chen, S. Zhang, and W. Lu, "Automated camera-exposure control for robust localization in varying illumination environments," *Autonomous Robots*, vol. 46, no. 4, pp. 515–534, 2022. 84

[258] Y. Yang, E. Rigall, H. Fan, and J. Dong, "Point light measurement and calibration for photometric stereo," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–11, 2023. 84

[259] P. Monkam, J. Wu, W. Lu, W. Shan, H. Chen, and Y. Zhai, "Easyspec: Automatic specular reflection detection and suppression from endoscopic images," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1031–1043, 2021. 85

[260] O. El Meslouhi, M. Kardouchi, H. Allali, T. Gadi, and Y. A. Benkaddour, "Automatic detection and inpainting of specular reflections for colposcopic images," *Central European Journal of Computer Science*, vol. 1, pp. 341–354, 2011. 85, 87

[261] S. Li, S. Zhu, Y. Ge, B. Zeng, M. A. Imran, Q. H. Abbasi, and J. Cooper, "Depth-guided deep video inpainting," *IEEE Transactions on Multimedia*, 2023. 86, 91

[262] J. Li, Y. Wen, and L. He, "Scconv: Spatial and channel reconstruction convolution for feature redundancy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6153–6162, 2023. 86

[263] P. E. Edwards, D. Psychogyios, S. Speidel, L. Maier-Hein, and D. Stoyanov, "Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction," *Medical Image Analysis*, vol. 76, p. 102302, 2022. 86, 96

[264] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012. 90

[265] X. Dray, A. Histace, A. Robertson, and S. Segui, "Artificial intelligence for protruding lesions," in *Artificial Intelligence in Capsule Endoscopy*, pp. 121–148, Elsevier, 2023. 91

[266] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020. 91

[267] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Free-form video inpainting with 3d gated convolution and temporal patchgan," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9066–9075, 2019. 91

[268] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711, Springer, 2016. 91

[269] P. Gómez, M. Semmler, A. Schützenberger, C. Bohr, and M. Döllinger, "Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network," *Medical & Biological Engineering & Computing*, vol. 57, pp. 1451–1463, 2019. 93

[270] J.-F. Pambrun and R. Noumeir, "Limitations of the ssim quality metric in the context of diagnostic imaging," in *2015 IEEE International Conference on Image Processing*, pp. 2960–2963, IEEE, 2015. 94

[271] T. Ben-Menachem, G. A. Decker, D. S. Early, J. Evans, R. D. Fanelli, D. A. Fisher, L. Fisher, N. Fukami, J. H. Hwang, S. O. Ikenberry, *et al.*, "Adverse events of upper gi endoscopy," *Gastrointestinal Endoscopy*, vol. 76, no. 4, pp. 707–718, 2012. 95

[272] H. Hu, Y. Chen, J. Xu, S. Borse, H. Cai, F. Porikli, and X. Wang, "Learning implicit feature alignment function for semantic segmentation," in *European Conference on Computer Vision*, pp. 487–505, Springer, 2022. 98

[273] X. Luo, K. Mori, and T. M. Peters, "Advanced endoscopic navigation: surgical big data, methodology, and applications," *Annual Review of Biomedical Engineering*, vol. 20, no. 1, pp. 221–251, 2018. 99

[274] J. Pottle, "Virtual reality and the transformation of medical education," *Future Healthcare Journal*, vol. 6, no. 3, pp. 181–185, 2019. 102

[275] R. Han, K. Zhou, A. Atapour-Abarghouei, X. Liang, and H. P. H. Shum, "Finecausal: A causal-based framework for interpretable fine-grained action quality assessment," in *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025. 102

[276] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, 2022. 103

[277] S. Rahman, S. Sarker, A. N. Haque, M. M. Uttsha, M. F. Islam, and S. Deb, "Ai-driven stroke rehabilitation systems and assessment: a systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 192–207, 2022. 103

[278] X. Zhang, F. Liang, C. T. Lau, J. C. Chan, N. Wang, J. Deng, J. Wang, Y. Ma, L. L. Zhong, C. Zhao, *et al.*, "Standards for reporting interventions in clinical trials of tuina/massage (strictotm): Extending the consort statement," *Journal of Evidence-Based Medicine*, vol. 16, no. 1, pp. 68–81, 2023. 103

[279] G. Karvounas, N. Kyriazis, I. Oikonomidis, and A. Argyros, "Dynamic multiview refinement of 3d hand datasets using differentiable ray tracing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3156–3166, 2023. 103

[280] J. E. Jang, Y. S. Lee, W. S. Jang, W. S. Sung, E.-J. Kim, S. D. Lee, K. H. Kim, and C. Y. Jung, "Trends in acupuncture training research: focus on practical phantom models," *Journal of Acupuncture Research*, vol. 39, no. 2, pp. 77–88, 2022. 103

[281] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2949–2958, 2022. 103

[282] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2405–2414, 2023. 103

[283] A. Al-Bedah, G. Ali, T. Abushanab, and N. Qureshi, "Tui na (or tuina) massage: a minireview of pertinent literature, 1970-2017," *Journal of Complementary and Alternative Medical Research*, vol. 3, no. 1, pp. 1–14, 2017. 104, 105

[284] G. Du, Y. Li, K. Su, C. Li, and P. X. Liu, "A mobile natural human–robot interaction method for virtual chinese acupuncture," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2022. 104, 109

[285] Resuscitation Council UK, "Quality standards: Introduction and overview," 2020. Published November 2013; updated June 2020. 104, 118

[286] W.-W. Tao, H. Jiang, X.-M. Tao, P. Jiang, L.-Y. Sha, and X.-C. Sun, "Effects of acupuncture, tuina, tai chi, qigong, and traditional chinese medicine five-element music therapy on symptom management and quality of life for cancer patients: a meta-analysis," *Journal of Pain and Symptom Management*, vol. 51, no. 4, pp. 728–747, 2016. 105

[287] S. Ardeshir and A. Borji, "An exocentric look at egocentric actions and vice versa," *Computer Vision and Image Understanding*, vol. 171, pp. 61–68, 2018. 106

[288] Q. Sun, J. Huang, H. Zhang, P. Craig, L. Yu, and E. G. Lim, "Design and development of a mixed reality acupuncture training system," in *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 265–275, IEEE, 2023. 106

[289] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357–366, 2021. 109

[290] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion*, pp. 1–6, IEEE, 2020. 110

[291] S. Han, P.-c. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan, *et al.*, "Umetrack: Unified multi-view end-to-end hand tracking for vr," in *SIGGRAPH Asia 2022 Conference*, pp. 1–9, 2022. 111

[292] J. Chen, R. Huang, K. Zhao, W. Wang, L. Liu, and W. Li, "Multiscale convolutional neural network with feature alignment for bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021. 112

[293] R. Lyu, M. Gao, H. Yang, Z. Wen, and W. Tang, "Stimulation parameters of manual acupuncture and their measurement," *Evidence-Based Complementary and Alternative Medicine*, vol. 2019, no. 1, p. 1725936, 2019. 115

[294] C. A. Smith, C. J. Zaslawski, S. Cochrane, X. Zhu, Z. Zheng, B. Loyeung, P. C. Meier, S. Walsh, C. C. Xue, A. L. Zhang, *et al.*, "Reliability of the nicman scale: an instrument to assess the quality of acupuncture administered in clinical trials," *Evidence-Based Complementary and Alternative Medicine*, vol. 2017, no. 1, p. 5694083, 2017. 115

[295] M. D. Constable, F. X. Zhang, T. Conner, D. Monk, J. Rajsic, C. Ford, L. J. Park, A. Platt, D. Porteous, L. Grierson, *et al.*, "Advancing healthcare practice and education via data sharing: demonstrating the utility of open data by training an artificial intelligence model to assess cardiopulmonary resuscitation skills," *Advances in Health Sciences Education*, pp. 1–21, 2024. 119

[296] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3d human pose estimation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404, 2019. 121

[297] Z.-h. Jian, M.-f. Sheng, J.-y. Li, D.-z. An, Z.-j. Weng, and G. Chen, "Developing a method to precisely locate the keypoint during craniotomy using the retrosigmoid keyhole approach: Surgical anatomy and technical nuances," *Frontiers in Surgery*, vol. 8, p. 700777, 2021. 126

[298] Z. Liu, H. Mu, X. Zhang, Z. Guo, X. Yang, K.-T. Cheng, and J. Sun, "Metapruning: Meta learning for automatic neural network channel pruning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3296–3305, 2019. 126

[299] A. Vettoruzzo, M.-R. Bouguelia, J. Vanschoren, T. Rognvaldsson, and K. Santosh, "Advances and challenges in meta-learning: A technical review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 126

[300] W. Du, A. Li, P. Zhou, B. Niu, and D. Wu, "Privacyeye: A privacy-preserving and computationally efficient deep learning-based mobile video analytics system," *IEEE Transactions on Mobile Computing*, vol. 21, no. 9, pp. 3263–3279, 2021. 126

[301] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024. 126

[302] H. Kassem, D. Alapatt, P. Mascagni, A. Karargyris, and N. Padoy, "Federated cycling (fedcy): Semi-supervised federated learning of surgical phases," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 1920–1931, 2022. 127

[303] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, "Challenges of deep learning in medical image analysis—improving explainability and trust," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 68–75, 2023. 127

[304] H. Zhang, *Human Movement Disorders Analysis with Graph Neural Networks*. PhD thesis, Durham University, 2024. 127

[305] H. Zhang, E. S. Ho, F. X. Zhang, and H. P. Shum, "Pose-based tremor classification for parkinson's disease diagnosis from video," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 489–499, Springer, 2022. 127

[306] Y. Liu, F. Liu, L. Jiao, Q. Bao, L. Li, Y. Guo, and P. Chen, "A knowledge-based hierarchical causal inference network for video action recognition," *IEEE Transactions on Multimedia*, 2024. 127

[307] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023. 127

[308] Y. Long, W. Wei, T. Huang, Y. Wang, and Q. Dou, "Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4441–4448, 2023. 128

[309] A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädsch, C. H. Sudre, L. Acion, M. Antonelli, *et al.*, "Understanding metric-related pitfalls in image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 182–194, 2024. 128

[310] R. Yagi, S. Goto, Y. Katsumata, C. A. MacRae, and R. C. Deo, "Importance of external validation and subgroup analysis of artificial intelligence in the detection of low ejection fraction from electrocardiograms," *European Heart Journal-Digital Health*, vol. 3, no. 4, pp. 654–657, 2022. 128

[311] R. Boulkedid, H. Abdoul, M. Loustau, O. Sibony, and C. Alberti, "Using and reporting the delphi method for selecting healthcare quality indicators: a systematic review," *PloS One*, vol. 6, no. 6, p. e20476, 2011. 128

[312] A. Domalpally and R. Channa, "Real-world validation of artificial intelligence algorithms for ophthalmic imaging," *The Lancet Digital Health*, vol. 3, no. 8, pp. e463–e464, 2021. 129

[313] D. Lin, J. Xiong, C. Liu, L. Zhao, Z. Li, S. Yu, X. Wu, Z. Ge, X. Hu, B. Wang, *et al.*, "Application of comprehensive artificial intelligence retinal expert (care) system: a national real-world evidence study," *The Lancet Digital Health*, vol. 3, no. 8, pp. e486–e495, 2021. 129

[314] S. Harris, T. Bonnici, T. Keen, W. Lilaonitkul, M. J. White, and N. Swanepoel, "Clinical deployment environments: Five pillars of translational machine learning for health," *Frontiers in Digital Health*, vol. 4, p. 939292, 2022. 129

APPENDIX A

Ethical Approvals

We provide the ethical approval supporting the publications included in this thesis.
Deep Learning-based Surgery Tools Motion Analysis (Reference ID: COMP-2022-10-26T13_23_46-slxb76, approved by the Ethics Committee of the Department of Computer Science at Durham University, UK, approval date: 22/01/2022)

Publications related to this approval:

- **Zhang, F. X.**, Moubayed, N. A., & Shum, H. P. H. (2022). Towards graph representation learning based surgical workflow anticipation. In *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI '22)*, pages 1-4, IEEE.

- **Zhang, F. X.**, Deng, J., Lieck, R., & Shum, H. P. H. (2025). Adaptive graph learning from spatial information for surgical workflow anticipation. *IEEE Transactions on Medical Robotics and Bionics, 7*(1), 266–280.

- **Zhang, F. X.**, Chen, S., Xie, X., & Shum, H. P. H. (2024). Depth-aware endoscopic video inpainting. In *Proceedings of the 2024 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '24)*, Springer, Marrakesh, Morocco, pp. 143–153.

Estimation for Health Professional Education: Development of an Objective Computerized Approach for Measuring and Assessing Technical Competencies in Nursing (Reference ID: COMP-2022-03-11T15_16_05-slxb76, approved by the Ethics Committee of the Department of Computer Science at Durham University, UK, approval date: 12/08/2022)

Publication related to this approval:

- Constable, M. D., **Zhang, F. X.**, Conner, T., Monk, D., Rajsic, J., Ford, C., Park, L. J., Platt, A., Porteous, D., Grierson, L., & Shum, H. P. H. (2024). Advancing Healthcare Practice and Education via Data Sharing: Demonstrating the Utility of Open Data by Training an Artificial Intelligence Model to Assess Cardiopulmonary Resuscitation Skills. In *Advances in Health Sciences Education*, pp. 1-21.

Egocentric Action Quality Assessment for Clinical Technique (Reference ID: COMP-2023-03-24T13_52_41-slxb76, approved by the Ethics Committee of the Department of Computer Science at Durham University, UK, approval date: 17/04/2023)

Publication related to this approval:

- **Zhang, F. X.**, Yao, H., Chen, S., Jia, X., Zheng, S., & Shum, H. P. H. (2025). Towards cross-view multimodality action quality assessment for Traditional Chinese Medicine physical therapy. Rejected with invitation to resubmit in March 2025; revision in preparation for *IEEE Transactions on Instrumentation and Measurement*.

APPENDIX B

Related Resources of Publications

We supplement the related resources (*e.g.*, paper, GitHub code, video, ArXiv, *etc.*) for all the publications by the author included in this thesis.

Zhang, F. X., Moubayed, N. A., & Shum, H. P. H. (2022). Towards graph representation learning based surgical workflow anticipation. In *Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI '22)*, pages 1-4, IEEE.

Related links: Paper GitHub Video ArXiv

Zhang, F. X., Deng, J., Lieck, R., & Shum, H. P. H. (2025). Adaptive graph learning from spatial information for surgical workflow anticipation. *IEEE Transactions on Medical Robotics and Bionics*, *7*(1), 266–280.

Related links: Paper GitHub Demo ArXiv

Zhang, F. X., Chen, S., Xie, X., & Shum, H. P. H. (2024). Depth-aware endoscopic video inpainting. In *Proceedings of the 2024 International Conference*

on Medical Image Computing and Computer Assisted Intervention (MICCAI '24), Springer, Marrakesh, Morocco, pp. 143–153.

Related links: Paper GitHub ArXiv

Constable, M. D., **Zhang, F. X.**, Conner, T., Monk, D., Rajsic, J., Ford, C., Park, L. J., Platt, A., Porteous, D., Grierson, L., & Shum, H. P. H. (2024). Advancing Healthcare Practice and Education via Data Sharing: Demonstrating the Utility of Open Data by Training an Artificial Intelligence Model to Assess Cardiopulmonary Resuscitation Skills. In *Advances in Health Sciences Education*, pp. 1-21.

Related links: Paper GitHub Dataset

Zhang, F. X., Yao, H., Chen, S., Jia, X., Zheng, S., & Shum, H. P. H. (2025). Towards cross-view multimodality action quality assessment for Traditional Chinese Medicine physical therapy. Rejected with invitation to resubmit in March 2025; revision in preparation for *IEEE Transactions on Instrumentation and Measurement*.

Related links: GitHub Dataset Demo

APPENDIX C

Other Publication

We supplement the related resources (*e.g.*, papers, GitHub code, ArXiv, *etc.*) for all other publications led by the author not included in this thesis but supervised during the author's PhD studies.

Zhang, F. X., Zheng, S., Shum, H. P. H., Zhang, H., Song, N., Song, M., & Jia, H. (2023). Correlation-Distance Graph Learning for Treatment Response Prediction from rs-fMRI. In *Proceedings of the 2023 International Conference on Neural Information Processing (ICONIP '23)* (pp. 298–312). Springer, Changsha, China. doi:10.1007/978-981-99-8138-0_24.

Related links: Paper GitHub ArXiv

Zheng, S., **Zhang, F. X.**, Shum, H. P. H., Zhang, H., Song, N., Song, M., & Jia, H. (2024). Unraveling the Brain Dynamics of Depersonalization-Derealization Disorder: A Dynamic Functional Network Connectivity Analysis. In *BMC Psychiatry*. BMC.

Related links: Paper