

Deep Learning Methods for Multi-level Surgical Video Scene Understanding

By

Luoying Hao

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
July 2025

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Medical video analysis spans a wide range of applications, offering valuable insights into patient care and clinical decision-making. Among these, surgical videos have emerged as an especially rich source of data, providing unique opportunities to analyze and optimize complex clinical workflows with high precision and contextual depth. This trend is particularly evident with the rise of Minimally Invasive Surgery (MIS). While MIS has transformed the operating room into a data-rich environment, it has also increased the complexity of surgical workflows and the cognitive workload on surgeons. Consequently, optimizing these workflows and reducing the burden on surgeons via intelligent systems that provide clinical decision support, known as Context-Aware Systems (CAS), is a growing necessity. A main component of such systems is the ability to automatically recognize the current state of the surgery. This is effectively accomplished by modeling workflows as a hierarchy of activities defined at different levels of detail, such as phases, steps, and actions. Despite extensive research in surgical activity recognition, a majority of these efforts have focused on coarse-grained phase recognition. A comprehensive, multi-level understanding remains a significant challenge, not only due to data scarcity but also due to key methodological limitations. Current models often fail to explicitly model the hierarchical context between activities, are susceptible to learning spurious correlations from visual biases, and lack the robustness to generalize across different clinical environments. Addressing these AI challenges is crucial for advancing the capabilities of CAS.

This thesis aims to develop advanced deep learning models for multi-level medical video scene understanding by tackling the aforementioned limitations. First, to address the lack

of suitable data, we construct a comprehensive cataract surgery dataset with synchronized, multi-level annotations. Next, we develop a novel detection network to improve the accuracy of fine-grained action recognition and, for the first time, to provide confidence scores for clinical trust. Subsequently, we propose a hierarchical framework that explicitly models the contextual relationships between activity levels. To tackle spurious correlations, we introduce a causality-inspired framework to mitigate confounding biases and enhance model robustness. Finally, we develop a foundation model and validate its generalization capabilities across multiple clinical centers.

Collectively, the findings demonstrate that by explicitly modeling hierarchical context and mitigating confounding biases through causality-inspired frameworks, robust and generalizable multi-level medical video understanding is achievable. The novel dataset and methodologies presented in this thesis provide a foundation for developing next-generation CAS, ultimately enhancing clinical training, safety, and efficiency.

CONTENT ADVISORY

This thesis contains clinical and surgical images that may be considered graphic or distressing to some readers. These images are included solely for scientific and educational purposes. Viewer discretion is advised.

ACKNOWLEDGMENTS

The years of my doctoral studies have passed in the blink of an eye. This journey has been filled with challenges, but it has also been a period of immense growth, bringing with it invaluable knowledge, experience, and friendships. The successful completion of this thesis would not have been possible without the care and support of many individuals, to whom I wish to express my most sincere gratitude.

As a joint Ph.D. student between the University of Birmingham (UoB) and Southern University of Science and Technology (SUSTech), I must first extend my special thanks to my supervisors: Professor Jinming Duan from UoB, and Professor Jiang Liu and Professor Yan Hu from SUSTech. Throughout my doctoral program, they not only provided patient academic guidance but also offered tremendous encouragement in my life and personal development. Whenever I encountered a research problem, they would help me clarify my thoughts and identify the core issues. During the writing of this thesis, their repeated reviews and invaluable suggestions were incredibly beneficial. Their rigorous approach to research and strong sense of responsibility have deeply influenced me and set an example I will always strive to follow.

I am grateful to my thesis examiners. Thank you for taking the time from your busy schedules to review my thesis and for providing precious recommendations that helped me further refine my research.

Furthermore, I extend heartfelt thanks to the respected members of my Thesis Group at the School of Computer Science at UoB: Professor Shan He (my second supervisor), Professor

Hyung Jin Chang, and Professor Jianbo Jiao. Your advice and guidance during the RSMG meetings were crucial to my research.

I would like to express my sincere gratitude to my research collaborators and fellow colleagues. At the UoB, I am very thankful for the support and friendship of Dr. Xi Jia, Dr. Tianyang Zhang, Yang Yue, Joseph Bartlett, Yuting Zhang, Xinxing Cheng, Tiantian Geng, Lingxi Hu and Mengjie Guo. I would also like to extend my thanks to my colleagues at SUSTech. I am grateful for the insightful discussions and help from Wenjun Lin, Jiaqi Zhang, Zhongxi Qiu, and Qinjie Jiang. To everyone, both mentioned and unmentioned, I offer my heartfelt thanks.

I am grateful to BEAR and Baskerville Tier 2 HPC services, operated by Advanced Research Computing at UoB, for providing the computational resources for part of my research. I also thank the iMED Lab at SUSTech for its generous support with funding, experimental equipment, and data.

This research was partially supported by the National Natural Science Foundation of China (Grant No. 82102189, 82272086). This funding was essential for the successful advancement of the project.

I owe a special debt of gratitude to my family. My parents have always been my strongest supporters, and their love and encouragement have been the greatest motivation for completing my studies. I also want to thank my partner, Mr. Wu. Thank you for your companionship throughout my postgraduate studies, for comforting and encouraging me when I felt lost, and for being a constant source of warmth and strength in my life.

Contents

Page

1 Introduction **1**

1.1 Surgical Data Science 5

1.2 Surgical Video Scene Understanding 7

1.2.1 Type of Activities 7

1.2.2 Challenges 10

1.2.3 Applications 16

1.3 Research Questions 17

1.4 Contribution of the Thesis 19

1.5 Publications 21

1.6 Thesis Structure 23

2 Related Works **25**

2.1 Surgical Activity Recognition 26

2.1.1 Coarse-grained Phase & Step Recognition 26

2.1.2 Fine-grained Action Recognition and Detection 29

2.1.3 Multi-task Learning in Surgical Scene 31

2.2 Causal Inference 32

2.2.1 Causal Inference in Medical Scene 35

2.3 Foundation Model in Medical Scene 36

2.4 Thesis positioning 39

3 Multi-level Surgical Video Dataset Construction **41**

3.1 Cataract Dataset 44

3.1.1 Ethical approval 44

3.1.2 Cataract Surgery 45

3.1.3 Label Definition 47

3.1.4 Dataset Statistics 47

3.2 Endoscopic Dataset 51

4 Fine-grained Surgical Video Scene Understanding **55**

4.1 Introduction 56

4.2 Methodology 58

4.2.1 Our Anchor-Context Detection Module 59

4.2.2 Conditional Diffusion for Action Detection 62

4.3 Experimental Results 65

4.3.1 Cataract Surgical Video Dataset 65

4.3.2 Implementation Details 65

4.3.3 Method Comparison 67

4.3.4 Ablation Study 67

4.3.5 Confidence Analysis 69

4.3.6 Qualitative Analysis and Limitations 72

4.4 Summary 73

5 Multi-level Surgical Video Scene Understanding **74**

5.1 Introduction 75

5.2 Related work 79

5.2.1 Surgical Scene Understanding 79

5.2.2 Multi-task Learning in Surgical Scene 80

5.3 Methodology 81

5.3.1 Overview 81
5.3.2 Hierarchical Relation Aggregation Module (HRAM) 83
5.3.3 Inter-task Contrastive Learning (ICL) 86
5.3.4 HCT+ Network 87
5.3.5 Overall Training Objective 89
5.4 Experiments 89
5.4.1 Datasets 89
5.4.2 Implementation Details 91
5.4.3 Comparison with State-of-the-arts 91
5.4.4 Ablation Study 98
5.4.5 Effectiveness of HCT+ 106
5.4.6 Result Visualisation and Analysis 109
5.4.7 Challenge Cases Analysis 115
5.5 Summary 116

6 Deconfounded Multi-level Surgical Video Scene Understanding with Causal Intervention 118

6.1 Introduction 119
6.2 Related work 123
6.2.1 Surgical Scene Understanding 123
6.2.2 Causal Inference 125
6.3 Methodology 125
6.3.1 Overview 125
6.3.2 The Video Transformer Architecture 127
6.3.3 Subject-Centric Video Representation 127
6.3.4 Surgical Activity Causal model 130
6.3.5 Overall Training Objective 136

6.4 Experiments . . . 136
6.4.1 Datasets . . . 136
6.4.2 Evaluation Metrics . . . 138
6.4.3 Implementation Details . . . 139
6.4.4 Comparison with state-of-the-arts . . . 139
6.4.5 Ablation Study . . . 148
6.4.6 Generalization Analysis . . . 157
6.4.7 Visualisation and Analysis . . . 160
6.5 Summary . . . 163

7 A Foundation Model for Multi-level Surgical Scene Understanding 165
7.1 Introduction . . . 166
7.2 Methodology . . . 168
7.2.1 Overview . . . 168
7.2.2 Spatial Vision Transformer (ViT) . . . 170
7.2.3 Temporal Vision Transformer (ViT) . . . 170
7.3 Experiments . . . 172
7.3.1 Cataract Datasets . . . 172
7.3.2 Implementation Details . . . 176
7.3.3 Evaluation Metrics . . . 178
7.3.4 Generalization Experiments . . . 178
7.4 Summary . . . 183

8 Conclusion and Future Work 189
8.1 Summary and Conclusion . . . 190
8.2 Future Work . . . 192

A Supplementary for Chapter 3 195

A.1 Label Definition for Cataract Dataset 196

References **206**

List of Figures

1.1 An example of a hybrid operating room. Image credit: [1] 2
1.2 Overview of Surgical Data Science (SDS) system. Image credit: [6]. 4
1.3 Surgical activities at various levels of granularity. 8
1.4 Surgical images illustrating several visual challenges: (a) Hand occlusion, (b) Specular reflection, (c) Blood splatter, (d) Presence of smoke, (e) Rapid motion blurring, (f) Out of body noise, (g) Restricted field-of-view, (h) Instrument occlusion. 13

2.1 Different methods for phase and step recognition. (a) LSTM-based method: SV-RCNet. Image credit: [91]. (b) TCN-based method: TeCNO. Image credit: [64]. (c) Transformer-based method: OperA. Image credit: [92]. . . . 27
2.2 Few examples of action triplets in laparoscopic cholecystectomy surgical videos. Image credit: [44]. 30
2.3 An example of causal graphs. Image credit: [113]. X: Treatment plays a direct causal role in determining the value of Outcome. Y: Outcome is the effect of the causal path. Z: Confounder is a common cause of both Treatment and Outcome. 33
2.4 Architecture of the Segment Anything Model (SAM). Image credit: [131]. . . 37

3.1 Cataract eye and cataract surgery illustration. (a) Comparison between a healthy eye and a cataract eye, highlighting the difference in lens status. Image credits: Healthdirect. (b) Diagram of cataract surgery, depicting the removal of the clouded lens and implantation of an artificial lens. Image credits: Guy's and St Thomas' Specialist Care [148]. 43
3.2 Surgical procedure demonstration based on real surgical data. Data source: Our cataract dataset. 46
3.3 Average duration of phases in the Cataract dataset. 48
3.4 Average duration of steps in the Cataract dataset. 48
3.5 Average duration of actions in the Cataract dataset. 49
3.6 Number of action types performed by different surgical instruments in the Cataract dataset. 49
3.7 Illustration of PSI-AVA dataset. Image credit: [21]. 51
3.8 PSI-AVA Statistics. Number of annotations for each class of the recognition and detection tasks. Colors denote the distribution in the fold partition. Image credit: [21]. 52

4.1 The pipeline of our Anchor-context Action Detection Network (ACTNet) includes Anchor-Context Detection (ACD) and Class Conditional Diffusion (CCD) modules. 59
4.2 Overview of the Class Conditional Diffusion (CCD) model. 61
4.3 Visualization on cataract dataset. We choose different actions to show the results of (a) Faster R-CNN, (b) OA_MIL, and (c) our ACTNet. For each example, we show the ground-truth (Blue), right predictions (Green) and wrong predictions (Red). The actions are labelled from left to right. IW values (multiplied by 100) mean prediction interval width to show the level of confidence. 68

4.4 Visualization on cataract dataset. We choose different actions to show the results of (a) Faster R-CNN, (b) OA_MIL, (c) our ACD and (d) our ACTNet. For each example, we show the ground-truth (Blue), right predictions (Green) and wrong predictions (Red). The actions are labelled from left to right. IW values indicate prediction Interval Width to show the level of confidence. . . 70

5.1 Comparison of our proposed method with existing methods. MSSU: Multi-level Semantic Scene Understanding in surgery. 76

5.2 The pipeline of our Hierarchical Context Transformer (HCT) framework. Given an input video with $l + 1$ frames, HCT uses a transformer model to extract shared features, which are then fed into the Hierarchical Relation Aggregation Module (HRAM) to capture the relations between the four task-wise features. After that, Inter-task Contractive Learning (ICL) is utilized to further optimize the HCT. For the plus version of the model HCT+, we add a temporal adapter before HRAM and put the spatial adapter in the feed-forward module. 82

5.3 Detailed structure of the Hierarchical Relation Aggregation Module (HRAM). 83

5.4 Detailed structure of the Inter-task Contrastive Learning (ICL). 85

5.5 Detailed structure of the spatial-temporal adapter in our proposed transformer block. 88

5.6 T-SNE Visualization results of (A) baseline, (B) baseline+HRAM and (C) HCT in phase, step and action recognition task. The blue circle indicates improvement in effectiveness. Due to space constraints, only the distribution of the 7 challenging action classes related to the instrument "Cannula" are shown in the figures, illustrating our method's enhanced performance on these difficult action classes. 100

5.7 Comparision of Average Precision (AP) of action detection across the challenging classes related to the instrument "Cannula" in Figure 5.6, to quantitatively demonstrate enhanced effectiveness. 101

5.8 Multi-level visualisation based on (A) Video66 from cataract dataset and (B) CASE015 from PSI-AVA dataset. (a) Qualitative comparisons for phase recognition. (b) Qualitative comparisons for step recognition. (c) Qualitative comparisons for action detection. We show the Ground-Truth (GT) (sky blue), right predictions (sky blue) and wrong predictions (light purple). 108

5.9 Qualitative comparison with our method and the state-of-the-art methods on two test video of cataract dataset for (A) phase recognition and (B) phase recognition. The color-coded ribbons illustrate the phases (P1-P4) and steps (S1-S11) during two videos with different durations. Horizontal axis represents the time progression. GT: Ground Truth. 111

5.10 Qualitative comparison with our method and the state-of-the-art methods on two test video of PSI-AVA dataset for (A) phase recognition and (B) phase recognition. The color-coded ribbons illustrate the phases (P1-P11) and steps (S1-S21) during two videos with different durations. Horizontal axis represents the time progression. GT: Ground Truth. 112

5.11 Multi-granularity visualisation based on Video30 from cataract dataset. (a) qualitative comparisons with the other six other methods for phase recognition. (b) qualitative comparisons with the other six other methods for step recognition. (c) qualitative comparisons with the other four methods for action detection. We show the ground-truth (sky blue), right predictions (sky blue) and wrong predictions (light purple). GT is the Ground Truth. 113

5.12 Multi-granularity visualisation based on CASE015 from PSI-AVA dataset. (a) qualitative comparisons with the other six methods for phase recognition. (b) qualitative comparisons with the other six methods for step recognition. (c) qualitative comparisons the other four methods for action detection.We show the ground-truth (sky blue), right predictions (sky blue) and wrong predictions (light purple). GT is the Ground Truth. 114

5.13 Color-coded ribbon visualization of phase and step recognition for some challenging classes from the proposed method (Pred) and the Ground Truth (GT) in two complete surgical videos. In each case, we show the classes with different colors and no class appearance with blank. Horizontal axes indicate the time progression. 115

6.1 Examples about the spurious correlation in surgical video understanding. We show the example from the action detection task in Cataract_ours. (a) shows the action concurrent rates with the instrument "irrigation aspiration handpiece". (b) shows the comparison of the t-SNE results for the transformer model and ours. In this case, the transformer learns the spurious correlation and struggles to separate two actions. Thanks to causal inference, ours achieves clearer clustering. 120

6.2 The pipeline of our Deconfounded Surgical Video Representation Learning (DSVRL) framework. Given an input video with $T + 1$ frames, DSVRL uses a transformer model and subject detector to extract corresponding features, which are then fed into the Subject-Centric Module (SCM) to capture subject-centric video representation. After that, the surgical Activity Causal Intervention Module (ACIM) is utilized to mitigate the effect of confounders, revealing the true causal relationships. The representations learned by the DSVRL framework can be utilized for three different granularities of video understanding tasks. 126

6.3 Detailed structure of the Subject-Centric Module (SCM). 128

6.4 Detailed structure of the Activity Causal Intervention Module (ACIM). . . . 130

6.5 Examples about the spurious correlation in step recognition. (a) shows the step concurrent rates with the instrument "irrigation handpiece". (b) shows a comparison of the qualitative results for the MViTv2 [211] and ours on the different frames. The predicted correct and incorrect results are colored green and red, respectively. Due to highly similar visual features and data bias, the MViTv2 [211] learns the spurious correlation in these cases and gives wrong results. Thanks to causal inference, ours achieves more accurate predictions. 131

6.6 Visualization comparison on few-sample classes to quantitatively demonstrate enhanced effectiveness. 148

6.7 T-SNE visualization results of (A) baseline, (B) baseline+SCM and (C) Ours in phase, step and action tasks. The blue circle indicates improvement in effectiveness. Due to space constraints, for action detection, only the distribution of the 11 challenging action classes related to the instrument "cannula" and "capsulorhexis forces" is shown in the figures, illustrating our method's enhanced performance on these difficult action classes. 151

6.8 Visualization of generalization experiments between Cataract_ours and CATARACTS on accuracy, precision, recall, and jaccard metrics. 159

6.9 Qualitative comparisons for action detection. Action detection locations of Ground Truth (GT) are indicated by green boxes, while the predicted action locations are shown with orange boxes. We also present the recognition results for GT (sky blue), right predictions (sky blue) and wrong predictions (light purple), along with the model's confidence. 161

6.10 Color-coded ribbon illustration for two complete surgical videos on step recognition. The time axes are scaled for better visualization. 162

6.11 Color-coded ribbon illustration for two complete surgical videos on phase recognition. The time axes are scaled for better visualization. 164

7.1 Workflow of this study. a. A pre-trained model: massive surgical videos from publicly available datasets and unlabeled videos are collected to pretrain the light Vision Transformer (ViT). b. Dataset Construction: multi-center ophthalmology surgical videos were converted to images for dataset construction. All frames from the datasets were annotated by annotators and reviewed by the expert surgeon. The annotation includes phase and step. c. Model Development: labeled surgical videos are used to train the COS model, which included pretrained ViT capturing spatial information and temporal ViT for temporal features among frames. d. Model evaluation: our COS model is tested on multi-center datasets. Its performance was evaluated using some metrics and feature/results visualization and compared with surgeon. COS Model: Comprehensive Ophthalmology Surgical Model. 169

7.2 Flowchart of our Comprehensive Ophthalmology Surgical (COS) Model for comprehensive ophthalmology surgical learning. The COS Model incorporates a two-step process for surgical learning in ophthalmology. In the first step, spatial features are extracted using the Vision Transformer (ViT) as the encoder, leveraging its pre-trained parameters. Recognizing the unique demands of surgical video analysis, we introduce a second step dedicated to temporal feature learning. This approach ensures a nuanced understanding of both spatial and temporal dimensions crucial for comprehensive surgical education. 171

7.3 The phase and step of a cataract surgery video are sampled in the figure. The instruments and image contrast are different from each center data. 175

7.4 Step/phase recognition for Cataract surgery. a. and b. The ROC was trained on the SZH dataset and tested on the SZH and WZH datasets, c. and d. The ROC trained on the publicly available Cataract101 dataset and tested on Cataract101 and WZH datasets. e. and f. The ablation study results prove the effectiveness of each component by phase and step recognition. g. The colour-coded ribbon illustration for a complete surgical video on step recognition. The COS model discovered outlier behaviors of two-times lens implant, which can indicates wrong operations, or complicated surgery. . . . 183

7.5 The ROC curves of phase recognition by different algorithms for cataract surgery. a. and b. are trained on SZH and test on SZH and WZH datasets. 185

7.6 The ROC curves of phase recognition by different algorithms for cataract surgery. a. and b. are trained on Cataract101 and test on Cataract101 and WZH datasets. 186

7.7 The ROC curves of step recognition by different algorithms for cataract surgery. a. and b. are trained on SZH and test on SZH and WZH datasets. 187

7.8 The ROC curves of step recognition by different algorithms for cataract surgery. a. and b. are trained on Cataract101 and test on Cataract101 and WZH datasets. 188

List of Tables

4.1 Methods comparison and ablation study on cataract video dataset. 66

4.2 The mean_IW (multiplied by 100) results from our CCD module. 72

5.1 Performance comparison on the cataract for phase and step recognition. The best results are highlighted in bold. "Frames" denotes the number of frames used in models. 92

5.2 Performance comparison on the cataract for action and tool detection. The best results are highlighted in bold. "Frames" denotes the number of frames used in models. 93

5.3 Performance comparison on the PSI-AVA for phase and step recognition. The best results are highlighted in bold. "Frames" denotes the number of frames used in models. 94

5.4 Performance comparison on the PSI-AVA for action and tool detection. The best results are highlighted in bold. "Frames" denotes the number of frames used in models. 95

5.5 Performance comparison on phase recognition based on BernBypass70 96

5.6 Performance comparison on step recognition based on BernBypass70 97

5.7 Performance Comparison on the Salad Dataset 98

5.8 Ablation study on key Components based on the two datasets. 99

5.9 Ablation study on HRAM. 102

5.10 Ablation study on ICL. The best results are highlighted in bold, while the next best results are in underline. 104
5.11 Performance comparison on new baseline model with the cataract dataset for four tasks 105
5.12 Performance comparison on new baseline model with the PSI-AVA for four tasks 106
5.13 Performance on the cataract dataset for HCT+. "Param" denotes the number of parameters and "PM" denotes peak memory usage. 107

6.1 Overview of four datasets 137
6.2 Performance comparison with other state-of-the-art methods on various datasets for action detection. The best results are highlighted in bold. 141
6.3 Performance comparison with other state-of-the-art methods on Cataract_ours dataset for step recognition. 142
6.4 Performance comparison with other state-of-the-art methods on CATARACTS dataset for step recognition. 142
6.5 Performance comparison with other state-of-the-art methods on GraSP dataset for step recognition. 143
6.6 Performance comparison with other state-of-the-art methods on Cataract_ours dataset for phase recognition. 145
6.7 Performance comparison with other state-of-the-art methods on CATARACTS dataset for phase recognition. 145
6.8 Performance comparison with other state-of-the-art methods on GraSP dataset for phase recognition. 146
6.9 Performance comparison with other state-of-the-art methods on AutoLaparo dataset for phase recognition. 147

6.10 Ablation study on key components for action detection, step recognition, and phase recognition on the Cataract_ours dataset. 150
6.11 Ablation study on attention structure. 153
6.12 Ablation study on confounder types. 153
6.13 Ablation study on cluster methods. 155
6.14 Ablation study on confounder sizes. 156
6.15 Ablation study on pre-trained datasets. 157
6.16 Comparison of generalization experiments between the Cataract_ours and CATARACTS. 158

7.1 The details of cataract surgery tasks and generalization to different hospitals. 176
7.2 Performance comparison: Train SZH → Test SZH (avg ± std) across all of the five folds. 179
7.3 Performance comparison: Train SZH → Test WZH (avg ± std) across all of the five folds. 179
7.4 Performance comparison: Train Cataract101 → Test Cataract101 (avg ± std) across all of the five folds. 180
7.5 Performance comparison: Train Cataract101 → Test WZH (avg ± std) across all of the five folds. 181

A.1 Definitions of all the proposed phases for the cataract surgery procedure. . . 196
A.2 Definitions of all the proposed steps for the cataract surgery procedure. . . . 197
A.3 Definitions of selected actions for the cataract surgery procedure. 199

Chapter One

Introduction



Figure 1.1: An example of a hybrid operating room. Image credit: [1]

.

Surgery, in its most fundamental definition, is a medical specialty that uses operative manual and instrumental techniques on a patient to investigate or treat a pathological condition such as a disease or injury, to help improve bodily function or appearance, or to repair unwanted ruptured areas. Historically, surgical interventions were often fraught with high risks and limited by the technology and understanding of the time. However, the relentless pursuit of improved patient outcomes has driven a remarkable evolution in surgical practices. Modern surgery is characterized by a multidisciplinary approach, integrating advanced imaging, sophisticated instrumentation, and a deeper understanding of physiology and pathology to offer increasingly precise and effective treatments [2, 3]. This progression has been significantly shaped by innovations aimed at reducing invasiveness, enhancing surgical capabilities, and leveraging the vast amounts of data generated within the surgical environment.

A significant development in this evolution has been the introduction of Minimally Invasive Surgery (MIS). Unlike traditional open surgery, which often requires large incisions, MIS techniques involve small incisions through which specialized instruments and a miniature camera (an endoscope or laparoscope) are inserted [4] This approach offers numerous patient benefits, including reduced postoperative pain, shorter hospital stays, faster recovery times, and improved cosmetic outcomes [5]. The expansion of MIS across various surgical disciplines, from cholecystectomies to complex cardiac procedures, has been facilitated by continuous advancements in instrument design, imaging technologies, and surgical training [4]. Despite its advantages, MIS presents unique challenges, such as a limited field of view, altered tactile feedback, and the need for specialized training to master the distinct hand-eye coordination required [5].

To address some of the specific challenges of MIS, such as a limited field of view, altered tactile feedback, and the need for specialized training to master the distinct hand-eye coordination required [5], and to further augment surgical capabilities, Computer-Assisted Intervention (CAI) emerged as a transformative force. CAI encompasses a broad range of technologies, including surgical navigation systems, robotic-assisted surgery, and advanced intraoperative imaging, all designed to enhance the surgeon's ability to plan, perform, and evaluate surgical procedures with greater accuracy and control [6]. Surgical robots, perhaps the most well-known facet of CAI, provide surgeons with enhanced dexterity, tremor filtration, and three-dimensional magnified vision, enabling more complex procedures to be performed minimally invasively [2]. These systems act as an extension of the surgeon's hands and eyes, translating their movements into precise actions at the surgical site. The integration of CAI has not only refined existing surgical techniques but has also enabled new therapeutic possibilities, pushing the boundaries of what can be achieved in the operating room [6]. Building upon the rich data streams generated by CAI systems and the increasingly digitized surgical environment, the field of Surgical Data Science (SDS) has recently

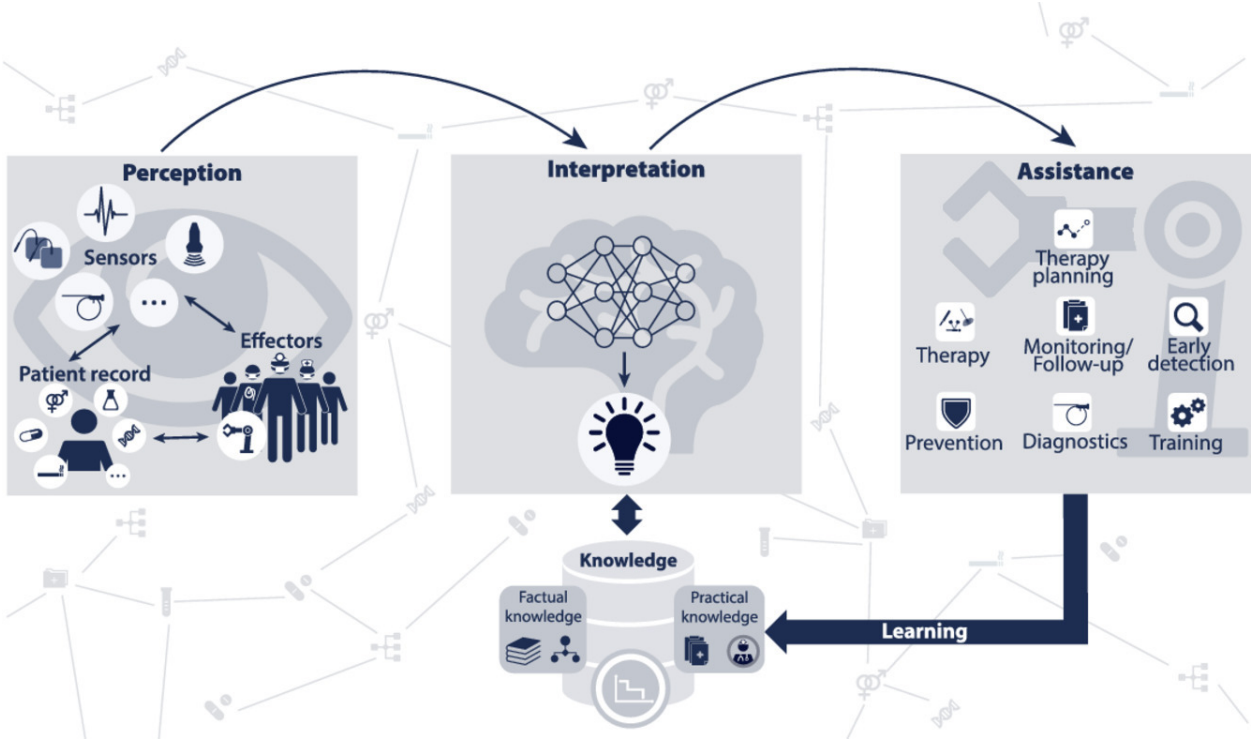


Figure 1.2: Overview of Surgical Data Science (SDS) system. Image credit: [6].

gained prominence [6].

The following sections present the aim of this thesis. First, the context of this study is introduced in Section 1.1, Surgical Data Science, and Section 1.2, Surgical Video Scene Understanding. These sections also discuss some of the fundamental research gaps and open questions in the community, which serve as the primary motivation for the research of this thesis. Next, Section 1.3 presents the core research questions this thesis aims to answer. Lastly, the contribution of the thesis is detailed in Section 1.4, the resulting publications are listed in Section 1.5, and the overall thesis structure is summarized in Section 1.6.

1.1 Surgical Data Science

Surgical Data Science (SDS) represents a crucial evolution in the quest to optimize surgical care, moving beyond the physical act of surgery to encompass the comprehensive analysis of all associated data [6, 7]. It leverages the increasing availability of digital information within the surgical pathway, from preoperative planning scans to intraoperative sensor readings and postoperative recovery metrics [8]. The core ambition of SDS is to transform this raw data into a structured understanding of surgical processes, which can then be used to develop intelligent systems capable of providing real-time support, personalized interventions, and continuous quality improvement [6, 9]. This involves developing methodologies for data acquisition, processing, and modeling, often employing sophisticated algorithms from machine learning and artificial intelligence to identify patterns, predict outcomes, and derive insights that might not be apparent through traditional observation alone [10]. Figure 1.2 provides an overview of the SDS system.

The relationship between CAI and Surgical Data Science is symbiotic and mutually reinforcing. CAI systems, particularly robotic platforms and advanced imaging modalities, are prolific generators of high-dimensional, time-synchronized data, including instrument kinematics, endoscopic video, and system logs [11]. This data serves as the primary fuel for SDS analyses. Conversely, the insights and models developed through SDS can be fed back to enhance CAI systems, making them smarter, more adaptive, and more attuned to the specific context of an ongoing procedure [12]. For instance, SDS can enable the development of algorithms that allow a surgical robot to anticipate a surgeon's next move or to provide warnings about potentially unsafe actions based on patterns learned from thousands of previous procedures.

A key application area for SDS, and a direct beneficiary of the CAI-SDS synergy, is the development of Context-Aware Systems (CAS) in surgery [12, 13]. CAS are intelligent systems

designed to perceive and interpret the current state of the surgical environment—including the patient's anatomy, the phase of the operation, the instruments in use, and the actions of the surgical team—to provide timely and relevant assistance [14, 15]. The goal of CAS is to enhance situational awareness, reduce cognitive load on the surgeon, improve decision-making, and ultimately, enhance patient safety and procedural efficiency [6, 12]. These systems integrate information from various sources, such as live video feeds, sensor data, and pre-operative plans, to build a comprehensive understanding of the surgical context [13]. This understanding allows the CAS to offer tailored support, such as displaying relevant anatomical information, highlighting critical structures, providing warnings about potential hazards, or even automating routine subtasks.

Central to the effective functioning of many CAS is the ability to automatically understand the content of surgical video streams. Surgical videos, typically captured via endoscopes or external cameras, provide a rich, dynamic visual record of the operative field [16, 17]. Automatic surgical video scene understanding involves the use of computer vision and machine learning techniques to analyze these videos in real-time to identify anatomical structures, recognize surgical instruments and their actions, determine the current phase of the surgery, and detect critical events [18, 19]. This capability is fundamental for CAS because it provides the system with a direct visual interpretation of what is happening at the surgical site, enabling it to react intelligently and provide contextually appropriate support [16]. For instance, by recognizing that the surgeon is about to perform a delicate dissection near a major blood vessel, a CAS could automatically adjust the display to provide a magnified view or overlay critical anatomical information. The accuracy and robustness of surgical video scene understanding algorithms are therefore paramount for the development of truly effective and reliable context-aware assistance in the operating room. To design an effective CAS, this thesis spotlights one of its key components, which is the automatic surgical video scene understanding.

1.2 Surgical Video Scene Understanding

Surgical Scene Understanding is a critical area within CAS that focuses on developing computational models capable of interpreting and analyzing activities occurring within the surgical environment. This thesis aims to implement a multi-level understanding of surgical scenarios, ranging from coarse-grained analysis, such as identifying broad surgical phases, to fine-grained analysis, which includes recognizing specific steps and elementary actions performed by the surgical team [20–22]. The overarching goal is to provide context-aware assistance, enhance surgical training, improve operating room efficiency, and ultimately contribute to better patient outcomes [6, 23]. By leveraging data from various sensors, primarily video feeds from endoscopic cameras or external room cameras, algorithms are developed to parse the complex and dynamic nature of surgical procedures into a structured and understandable format [24].

1.2.1 Type of Activities

Surgical procedures can be deconstructed into a hierarchy of activities, often categorized into phases, steps, and actions [25, 26] (Figure 1.3). This hierarchical decomposition allows for a structured analysis of the surgical workflow, where each level provides a different degree of detail about the ongoing procedure [25]. Phases represent broad segments of the surgery, steps are more specific tasks within those phases, and actions are the most granular movements or maneuvers performed by the surgeon [26]. The definition and delineation of these activities are typically derived from established surgical protocols, expert consensus among surgeons, and detailed observation of surgical procedures [27]. This structured approach facilitates a common language and framework for describing and analyzing surgical interventions [26, 28].

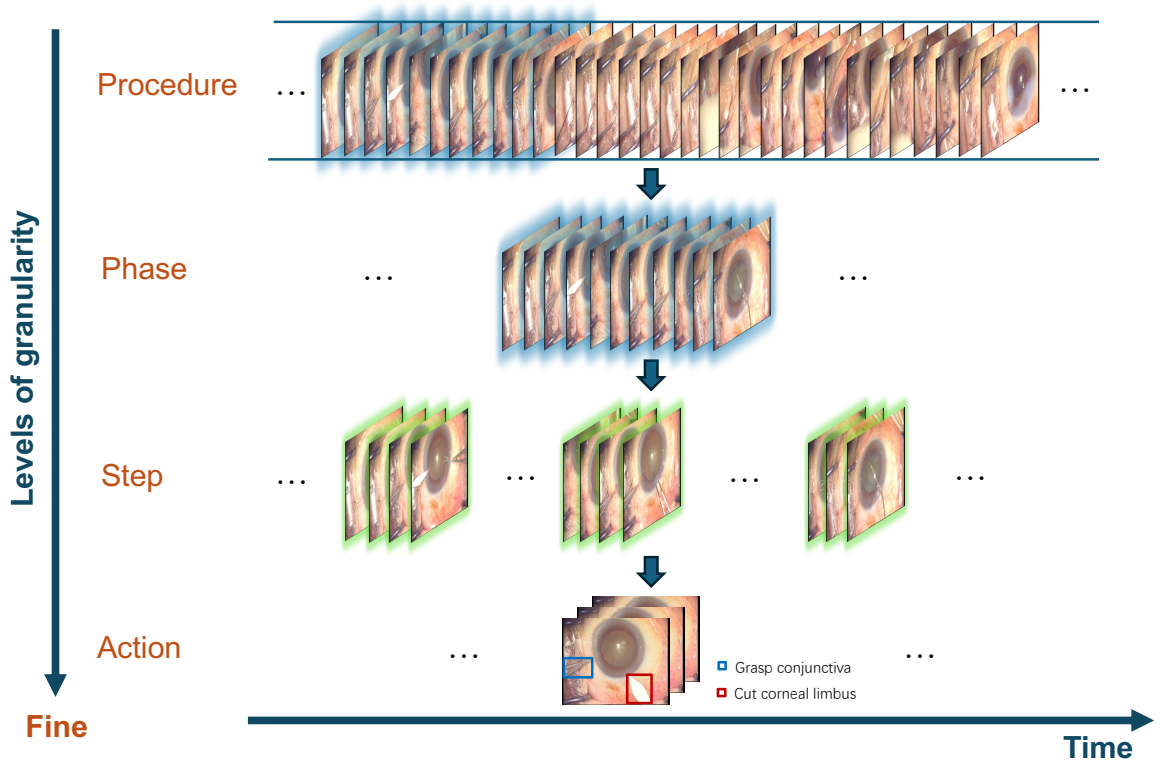


Figure 1.3: Surgical activities at various levels of granularity.

1.2.1.1 Phase

A surgical phase is defined as a major, distinct, and temporally ordered segment of an operation, representing a high-level surgical objective or a significant portion of the procedure [25]. Phases describe the surgical workflow at the coarsest level of granularity and typically occur in a sequential, though sometimes flexible, order to achieve the overall goals of the surgery [26]. The origin of phase definitions often stems from standardized surgical training, established clinical pathways, and expert consensus that divides a complex procedure into logical, manageable parts [29]. For example, in cataract surgery, common phases might include "Preparation and Incision", "Capsulorhexis", "Phacoemulsification", "Irrigation/Aspiration" and "Intraocular Lens (IOL) Implantation" [30, 31]. Each phase encompasses a series of more detailed steps required to complete that particular stage of the operation and has a clear start and end point, often marked by specific surgical milestones

[25, 32]]. The relationship between phases is generally sequential, forming the main timeline of the surgical intervention [33].

1.2.1.2 Step

A surgical step is a more fine-grained component of a surgical phase, representing a specific, clinically meaningful task or a sequence of activities performed to achieve a particular objective within that phase [34]. Steps are the intermediate level in the surgical activity hierarchy, breaking down the broader aims of a phase into more detailed procedural components [35]. The definition of surgical steps is typically derived from detailed procedural descriptions found in surgical textbooks, training manuals, and through direct observation and consensus among experienced surgeons who identify key operative tasks necessary to complete each phase [27]. For instance, within the "Phacoemulsification" phase of cataract surgery, distinct steps could include "Grooving/Sculpting the Nucleus", "Cracking the Nucleus" and "Emulsification and Aspiration of Nuclear Fragments" [30, 31]. While steps are generally performed to complete a phase, their exact order can sometimes vary, and some steps might be repeated or omitted based on the specific circumstances of the surgery [36]. Each step is composed of multiple, even more granular, surgical actions [34, 35].

1.2.1.3 Action

A surgical action, sometimes referred to as a gesture or atomic action, represents the most fine-grained, fundamental physical movement or maneuver performed by the surgeon, often involving the manipulation of a specific surgical instrument to achieve an immediate, low-level goal [34]. Actions are the building blocks of surgical steps and constitute the lowest level in the hierarchy of surgical activities [35, 37]. The origin of action definitions usually comes from detailed ergonomic studies of surgical movements, micro-analysis of surgical video, and

expert surgeons' decomposition of tasks into their elementary components [37]. In cataract surgery, during the "Emulsification and Aspiration of Nuclear Fragments" step, examples of actions could include "inserting phaco probe", "applying ultrasonic power", "aspirating cortical material" or "retracting instrument" [38]. These actions, when combined in specific sequences, form a surgical step, which in turn contributes to the completion of a surgical phase [25, 26]. The precise and coordinated execution of these actions is fundamental to the successful performance of the overall surgical procedure [37].

1.2.2 Challenges

Although researchers have shown significant interest in recognizing surgical activities, the majority of work has focused primarily on phase recognition. Both action detection and multi-level scene understanding remain underdeveloped areas, largely hindered by limitations in available datasets and methodological approaches.

1.2.2.1 Challenges Related to Data

Multi-level surgical scene understanding is crucial for the development of CAS. However, progress is often hindered by significant data-related challenges. This section outlines two primary obstacles: the scarcity of comprehensive multi-level activity datasets and the inherent visual complexities within surgical images.

- **Scarcity of Multi-level Activity Datasets:** Current research in surgical activity recognition has predominantly focused on individual tasks such as identifying surgical phases (coarse-grained segments of a procedure), recognizing steps (intermediate-level actions), or detecting specific actions and instrument usage (fine-grained activities) [20, 39]. While valuable, these studies on isolated tasks often rely on large-scale datasets with annotations tai-

lored to a single type of activity [33, 39]. For instance, the Cholec80 dataset is widely used for phase recognition in cholecystectomy procedures, providing phase and tool presence annotations for 80 videos [33, 40]. Similarly, the CATARACTS dataset, with its 50 videos, has been instrumental in surgical tool presence detection and activity recognition in cataract surgeries [41, 42]. More recent datasets like Cataract-1K offer a larger scale with 1000 cataract surgery videos, annotated for phases in a subset of 56 videos, along with some pixel-level annotations for instruments and anatomy [17].

The HeiChole dataset, used in the 2019 Endoscopic Vision challenge, provided 33 laparoscopic cholecystectomy videos annotated for phases, actions, instruments, and skills, highlighting the complexity of multi-center data [43]. CholecT50 was introduced to foster research in fine-grained activity recognition using surgical action triplets (instrument-verb-target) from 50 endoscopic videos of cholecystectomy, also including phase labels [44]. The MISAW sub-challenge as part of MICCAI 2020 focused on multi-granularity recognition (phases, steps, activities) using stereoscopic video and kinematic data from micro-surgical anastomosis training [45].

Despite these efforts, studies investigating integrated, multi-level surgical activity recognition—encompassing phases, steps, and actions simultaneously—remain scarce [20, 21]. Existing single-task datasets, while significant for their specific purposes, have limitations for developing holistic surgical scene understanding models. They often lack the hierarchical and concurrent annotations across multiple levels of granularity that are necessary to capture the full complexity of surgical procedures [20, 46]. This scarcity means that models trained on these datasets may not effectively learn the interdependencies between different levels of surgical activity.

The innovation and necessity of this thesis's attempt to construct a multi-level surgical video dataset lie in addressing this critical gap. A dataset with comprehensive, synchronized

annotations for phases, steps, and fine-grained actions within the same procedures will enable the development and validation of more sophisticated models. Such models can achieve a deeper understanding of the surgical workflow, leading to more effective context-aware assistance, improved surgical training, and enhanced postoperative analysis. The MultiBypass140 dataset, focusing on phase and step recognition in Roux-en-Y gastric bypass surgery, is a step in this direction, but a broader effort covering more granular details across various procedures is still needed [47].

- **Visual Challenges in Surgical Images:** Surgical videos present numerous visual challenges that can significantly impede automated analysis and understanding of the surgical scene. These challenges can degrade the performance of computer vision algorithms, making tasks like instrument tracking, tissue identification, and activity recognition more difficult [48, 49].

Occlusion is a pervasive issue, where instruments, hands, or anatomical structures obscure the surgical site of interest [50]. This partial visibility can lead to failures in tracking instrument poses and in accurately delineating anatomical boundaries [50]. Compounding this is the problem of specular reflection, where intense light from the endoscope reflects off the wet, smooth surfaces of tissues and the metallic surfaces of instruments, creating bright highlights [51]. These reflections can saturate image sensors, erasing crucial texture and color information, which interferes with feature extraction and can be mistaken for distinct objects by computer vision models [51].

The surgical process itself introduces further visual confounders. The use of electrosurgical tools frequently generates smoke, which can obscure large portions of the field of view, reducing visibility and contrast [52, 53]. Similarly, blood splatter and bodily fluids can cover the camera lens or coat tissues, altering their appearance and concealing underlying structures [52, 54]. These elements introduce significant noise and artifacts that require

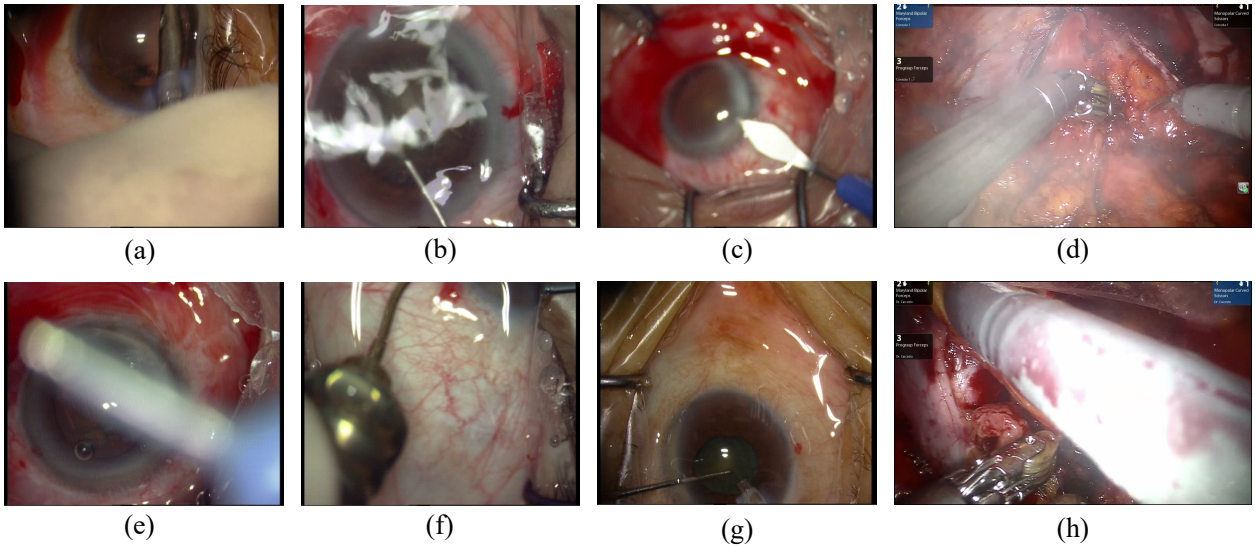


Figure 1.4: Surgical images illustrating several visual challenges: (a) Hand occlusion, (b) Specular reflection, (c) Blood splatter, (d) Presence of smoke, (e) Rapid motion blurring, (f) Out of body noise, (g) Restricted field-of-view, (h) Instrument occlusion.

sophisticated mitigation strategies to ensure reliable scene interpretation [55].

Furthermore, challenges are inherent to the image acquisition process in minimally invasive surgery. Rapid motion blurring occurs due to swift movements of the endoscope or instruments, resulting in indistinct frames that are difficult to analyze [16, 56]. The camera itself offers a restricted field-of-view, providing only a limited perspective of the entire surgical field [57, 58]. This narrow viewpoint means that critical context may be located just outside the frame, complicating the understanding of the broader surgical procedure and spatial relationships between tools and anatomy [57, 59]. Finally, variations in illumination and out-of-body noise, such as camera adjustments, further contribute to the complexity of the visual data stream [56, 58]. These visual challenges, as illustrated in Figure 1.4, make it difficult to design discriminative visual features to represent the data. Addressing these multifaceted visual challenges is a foundational prerequisite for developing intelligent systems capable of comprehensive surgical scene understanding.

1.2.2.2 Challenges Related to Methods

Methodological challenges significantly impede the advancement of surgical scene understanding. These challenges stem from the inherent complexity of surgical procedures and the limitations of current analytical approaches.

Firstly, the task of fine-grained surgical action detection is fraught with complexity. A single surgical instrument is often used to perform a wide variety of actions, and conversely, a single action may involve multiple instruments. Furthermore, surgical scenes are dynamic, with two or more distinct actions frequently occurring simultaneously. This concurrency presents a significant challenge for models, which must accurately detect the precise spatio-temporal location of each action while also correctly identifying its class from a large number of possibilities [60]. Current research in this domain remains limited, with most studies focusing on the coarser level of surgical phase recognition [22, 61]. Existing fine-grained action detection models often yields modest accuracy [39]. A critical deficiency in the current methodological landscape is the lack of robust analysis of prediction uncertainty [62, 63]. Without a reliable measure of a model's confidence in its predictions, it is difficult to assess the trustworthiness of the output, a crucial factor for adoption in safety-critical clinical environments [62]. Developing methods that can not only recognize complex, overlapping actions but also quantify the certainty of those detections is therefore a key challenge for advancing the field.

A second major challenge is the lack of a holistic, multi-level understanding of the surgical scene. The majority of existing work addresses surgical workflow analysis by targeting a single, isolated task, such as recognizing phases, steps, or actions independently [20, 64]. This approach fails to capitalize on the inherent hierarchical and contextual relationships between different levels of surgical activity. A surgical procedure is a structured process where high-level phases are composed of sequential steps, which in turn are made up of fine-grained

actions [20]. Developing integrated frameworks that can simultaneously model and reason across these multiple granularities is essential for building a comprehensive understanding of the surgical process, but such systems remain underdeveloped [7, 65].

Furthermore, many current methodologies overlook the exploration of essential causal relationships within the surgical workflow. Models often learn spurious correlations from the training data—for example, associating an instrument with an action simply because they frequently co-occur in the dataset, rather than learning the true causal link between them [66, 67]. This reliance on statistical shortcuts leads to a negative impact on performance, particularly for rare but clinically significant event classes which are underrepresented in training data. The presence of confounding factors can severely degrade a model's ability to understand the surgical scene, highlighting the need for methods that can move beyond simple pattern recognition to infer causality [68].

Finally, a persistent challenge is the limited generalization of proposed methods across different clinical environments. Models for surgical activity recognition are often trained and evaluated on data from a single institution or a specific surgical procedure [47]. However, these models frequently experience a significant drop in performance when applied to data from a different hospital, surgeon, or even a different operating room, a phenomenon known as domain shift [69]. This lack of robustness and poor transferability is a major obstacle to the widespread clinical implementation of surgical AI [66]. Consequently, developing domain generalization techniques that enable models to maintain high performance across diverse datasets and centers is a critical area of ongoing research [70].

1.2.3 Applications

Surgical video scene understanding, the automated interpretation of activities and events within the surgical field from video data, holds transformative potential across various facets of surgical practice. Its applications aim to enhance patient safety, improve training efficacy, optimize operative workflows, and foster continuous quality improvement. By converting vast streams of visual data into actionable insights, these technologies are poised to become indispensable tools in modern surgery.

- **Enhanced Intraoperative Decision Support and Context-Aware Assistance:** Surgical video understanding can provide real-time, context-relevant information to the surgical team. This includes highlighting critical anatomical structures, predicting surgeon's next actions, or warning about potential hazards based on the current surgical phase and instrument movements [49, 71]. For instance, by automatically recognizing that a surgeon is approaching a major blood vessel, the system could issue an alert or overlay relevant anatomical data onto the display. This enhanced situational awareness can reduce cognitive load and support safer, more precise interventions [71]. Key technical requirements include low-latency video processing and highly accurate recognition algorithms robust to visual challenges like smoke or bleeding [72].

- **Objective Surgical Skill Assessment and Personalized Training:** Automated analysis of surgical videos offers an objective and scalable method for assessing surgical skill [73, 74]. By identifying surgical phases, steps, instrument movements, and efficiency of motion, these systems can provide detailed, data-driven feedback to trainees and experienced surgeons alike [75, 76]. For example, AI models can compare a trainee's performance against expert benchmarks, identifying areas for improvement in technique or decision-making [77, 78]. This facilitates personalized training pathways and can accelerate the learning curve [75, 79]. The development of robust metrics for skill and clear visualization of performance

are crucial for this application [80].

• **Optimization of Operating Room Workflow and Efficiency:** Understanding the progression of a surgical procedure through video analysis can lead to significant improvements in Operating Room (OR) efficiency [81]. By automatically recognizing surgical phases and predicting remaining procedure times, systems can help optimize OR scheduling, resource allocation, and team coordination [81, 82]. For instance, accurately predicting the end of a surgery allows for better preparation for the next case, reducing downtime [81]. This requires integration with hospital information systems and robust models that can account for procedural variability [82, 83].

• **Postoperative Analysis, Quality Improvement, and Research:** Surgical video archives, when analyzed at scale, can become a rich resource for postoperative analysis, quality improvement initiatives, and surgical research [77, 84]. By correlating video-derived metrics (e.g., adherence to critical safety steps, occurrence of adverse events) with patient outcomes, researchers can identify best practices and areas for systematic improvement [77]. For example, analyzing videos of procedures with complications might reveal patterns or deviations from standard techniques that contributed to the adverse outcome [71]. This necessitates large, curated, and well-annotated video datasets and powerful analytical tools to extract meaningful correlations [85].

1.3 Research Questions

This thesis is focused on how to achieve effective multi-level understanding of surgical scenes. To reach this goal, several main research questions are proposed:

1. **How can we build a comprehensive dataset for multi-level surgical scene**

understanding?

For accurate multi-level understanding, it is important to construct datasets that include detailed and synchronized labels for surgical phases, steps, and fine-grained actions. This research investigates how such datasets can be created to support the development of robust models.

2. **How can we improve the detection of fine-grained surgical actions?**

   Recognizing small and subtle actions in surgery is a challenging but essential part of multi-level understanding. This thesis explores new methods for accurately detecting these fine-grained actions, and studies how these improvements can help in recognizing higher-level steps and phases.

3. **How can we make use of the interactions between different levels of surgical activities?**

   A truly effective framework should leverage the relationships between phases, steps, and actions. This work examines how to model and use these interactions to enhance the overall understanding of the surgical workflow.

4. **How can we explore and use causal relationships in surgical scene understanding?**

   Beyond finding patterns, it is important to identify the real causal links between surgical data (like tool movements or visual cues) and surgical events or predictions. This thesis looks into how these causal relationships can be incorporated to make the understanding framework more reliable and explainable.

5. **How can we improve generalization across different datasets and centers?**

   Finally, this research addresses how to ensure that the developed methods perform well not only on one dataset or in one medical center, but also across different datasets and institutions.

By addressing these questions, this thesis aims to provide new insights and practical solutions for multi-level surgical scene understanding.

1.4 Contribution of the Thesis

The fundamental aim of this thesis is to address the problem of surgical scene understanding at different levels of granularity from surgical videos, such as those from endoscopic and cataract procedures, by developing advanced deep learning models. We first construct a multi-level cataract surgery dataset for model development and validation. The method contributions are discussed in detail as follows:

The first contribution focuses on the critical task of recognizing and localizing detailed surgical actions, which is essential for developing context-aware decision support systems. Recognizing that most existing detection algorithms struggle with high-accuracy action classification despite locating actions, due to not considering the procedural regularity, and noting the unexplored need for model confidence in clinical applications, this work introduces an Anchor-context Action Detection Network (ACTNet). ACTNet comprises an Anchor-Context Detection (ACD) module to ascertain where actions occur and what they are by highlighting spatio-temporal regions interacting with extracted anchors, and a Class Conditional Diffusion (CCD) module to refine action predictions by leveraging a denoising diffusion-based generative model. Crucially, this approach also utilizes the stochastic nature of the diffusion model to provide confidence scores for each prediction. This method achieved state-of-the-art performance, demonstrating a significant 4.0% mAP improvement against baseline on our cataract video dataset.

As a second contribution, the thesis addresses the need for a systematic, hierarchical understanding of surgical scenes to enable advanced computer-assisted systems. Few existing

works [21] provide a comprehensive analysis across multiple levels of surgical activity. To this end, we propose a Multi-level Semantic Scene Understanding (MSSU) framework, conceptualizing the tasks of phase recognition, step recognition, and action and instrument detection as an interconnected hierarchy. A novel Hierarchical Context Transformer (HCT) network was developed, featuring a Hierarchical Relation Aggregation Module (HRAM) to concurrently model interactions within multi-level information and augment task-specific features. To further enhance representation learning, Inter-task Contrastive Learning (ICL) was introduced to guide the model in learning task-wise features by absorbing complementary information from other tasks. Additionally, a computationally efficient variant, HCT+, was proposed, integrating spatial and temporal adapters to achieve competitive performance with substantially fewer tunable parameters. Extensive experiments on our proprietary cataract dataset and the publicly available endoscopic PSI-AVA dataset [21] demonstrated the outstanding performance of this method, consistently exceeding state-of-the-art approaches by a large margin.

For the third contribution, this work tackles a long-overlooked issue in surgical scene understanding: the detrimental impact of hidden visual confounders, which can lead to harmful bias, induce spurious correlations, and degrade model generalization across different tasks and granularities [86]. To address the challenge of designing a robust, unified spatio-temporal representation that is resilient to such biases, we propose a causality-inspired framework for surgical video representation learning. This framework aims to capture the true causal effects of video content on predictions. Specifically, a Subject-Centric Module (SCM) is introduced within the transformer block to highlight critical regions by emphasizing the role of instrument subjects, leading to efficient and compact representations. Furthermore, a surgical Activity Causal Intervention Module (ACIM), based on backdoor adjustment derived from a tailored causal graph, is proposed to disentangle the model from the impact of harmful bias and exploit true causal effects for model training. This ensures robust and unbiased

representation learning. The effectiveness of this method and the significance of causal insight were validated through extensive experiments across four datasets (Cataract_ours, CATARACTS [41], GraSP [20], and AutoLaparo [87]) and three tasks (phase recognition, step recognition, and action detection), with cross-dataset experiments clearly demonstrating the model's superior generalization capabilities.

As a final contribution, we focus our attention on developing and validating a comprehensive foundation model for multi-level workflow recognition in cataract surgery, specifically designed for robust generalization across multiple clinical centers. Unlike previous approaches that are often limited to single-center data or narrow workflow segmentation, the proposed model leverages hierarchical spatiotemporal feature extraction to accurately identify surgical phases and steps across diverse environments, operator techniques, and patient populations. Extensive multi-center evaluations demonstrate the model's ability to reliably adapt to inter-center variability, establishing a new benchmark for scalable, standardized workflow analysis in cataract surgery. Our results highlight the potential of foundation models to advance the unify and optimize surgical workflow analysis.

1.5 Publications

This thesis has led to the following peer-reviewed papers:

- **Hao, L.**, Hu, Y., Lin, W., Wang, Q., Li, H., Fu, H., Duan, J. & Liu, J. (2023). Actnet: Anchor-context action detection in surgery videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 196-206). Cham: Springer Nature Switzerland.

- **Hao, L.**, Hu, Y., Yue, Y., Wu, L., Fu, H., Duan, J., & Liu, J. (2025). Hierarchical Con-

text Transformer for Multi-Level Semantic Scene Understanding. IEEE Transactions on Circuits and Systems for Video Technology, 35(4), 3134-3145.

- **Hao, L.**, Hu, Y., Dong, B., Fu, H., Duan, J., & Liu, J. (2025). Deconfounded Surgical Video Representation Learning with Causal Intervention. IEEE Transactions on Medical Imaging, Under Review.

- Hu, Y., **Hao, L.**, et al. (2025). Comprehensive Ophthalmology Surgical Foundation Model for Surgeon Activity Recognition and Evaluation. npj Digital Medicine, Co-first Author, Under Review.

These four papers correspond to Chapters 4, 5, 6, and 7 of this thesis. Other publications I made contributions during my PhD are listed as follows:

- **Hao, L.**, Hu, Y., Xu, Y., Fu, H., Miao, H., Zheng, C., & Liu, J. (2022). Dynamic analysis of iris changes and a deep learning system for automated angle-closure classification based on AS-OCT videos. Eye and Vision, 9(1), 41.

- **Hao, L.**, Hu, Y., Higashita, R., JQ Yu, J., Zheng, C., & Liu, J. (2022). Multiview volume and temporal difference network for angle-closure glaucoma screening from AS-OCT videos. Journal of Healthcare Engineering, 2022(1), 2722608.

- Lin, W., Hu, Y., **Hao, L.**, Zhou, D., Yang, M., Fu, H., Chui, C. & Liu, J. (2022). Instrument-tissue interaction quintuple detection in surgery videos. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 399-409). Cham: Springer Nature Switzerland.

- Zhang, Q., **Hao, L.**, Li, H., Ren, Z., & Cheng, J. (2022). GAN-SLAM: GAN based monocular visual-inertial simultaneous localization and mapping in dark environments. In 2022 5th International Symposium on Autonomous Systems (pp. 1-6). IEEE.

- Zhou, X., **Hao, L.**, Nie, Q., Zhou, Y., Wang, L., Hu, Y., & Liu, J. (2022). A novel multi-focus fusion network for retinal microsurgery. In 2022 IEEE 19th International Symposium on Biomedical Imaging (pp. 1-5). IEEE.

- Yue, Y., McDonald, D., **Hao, L.**, Lei, H., Butler, M. S., & He, S. (2023). FLONE: fully Lorentz network embedding for inferring novel drug targets. Bioinformatics Advances, 3(1), vbad066.

- Yue, Y., Liu, Y., **Hao, L.**, Lei, H., & He, S. (2023). Improving therapeutic synergy score predictions with adverse effects using multi-task heterogeneous network learning. Briefings in Bioinformatics, 24(1), bbac564.

- Lin, W., Hu, Y., **Hao, L.**, Fu, H., Chui, C. & Liu, J. (2025). AIPNet: Action-Instance Progressive Learning Network for Instrument-tissue Interaction Detection. IEEE Journal of Biomedical and Health Informatics.

1.6 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 presents a comprehensive review of the related works fundamental to this research. It covers surgical activity recognition, including both coarse-grained and fine-grained approaches, multi-task learning in surgical scenes, and the role of causal inference and foundational models in medical science.

Chapter 3 describes the construction of the multi-level surgical video dataset used throughout this thesis. This chapter details the process of data collection in cataract surgery, the definition of annotation labels, and an analysis of dataset statistics, providing the foundation for subsequent experimental studies.

Chapter 4 focuses on fine-grained surgical video scene understanding. It introduces the proposed methodologies for action detection, including the Anchor-Context Detection (ACD) module and conditional diffusion techniques. The chapter further presents experimental results, such as method comparisons, ablation studies, and confidence analysis on the cataract surgical video dataset.

Chapter 5 extends the discussion to multi-level surgical video scene understanding. It reviews related works on surgical scene understanding and multi-task learning, introduces novel methodologies such as the Hierarchical Relation Aggregation Module (HRAM) and Inter-task Contrastive Learning (ICL), and presents the HCT+ network. Comprehensive experiments, including comparisons with state-of-the-art methods, ablation studies, and analysis of challenging cases, are provided.

Chapter 6 addresses the challenge of deconfounded multi-level surgical video scene understanding. This chapter introduces causal inference into the video understanding pipeline, describing the video transformer architecture, subject-centric video representation, and a surgical activity causal model. Extensive experiments are conducted to evaluate the effectiveness and generalization of the proposed approaches.

Chapter 7 addresses the critical challenge of model generalization across different clinical environments. This methodology, based on spatial and temporal Vision Transformers (ViT), is specifically designed to learn robust features that mitigate domain shift. The chapter provides extensive experimental evaluation on multi-center cataract datasets, implementation details, performance metrics, and a primary focus on generalization experiments to validate the model's transferability.

Chapter 8 concludes the thesis by summarizing the key findings and contributions, discussing their broader implications, and providing directions for future research.

Chapter Two

Related Works

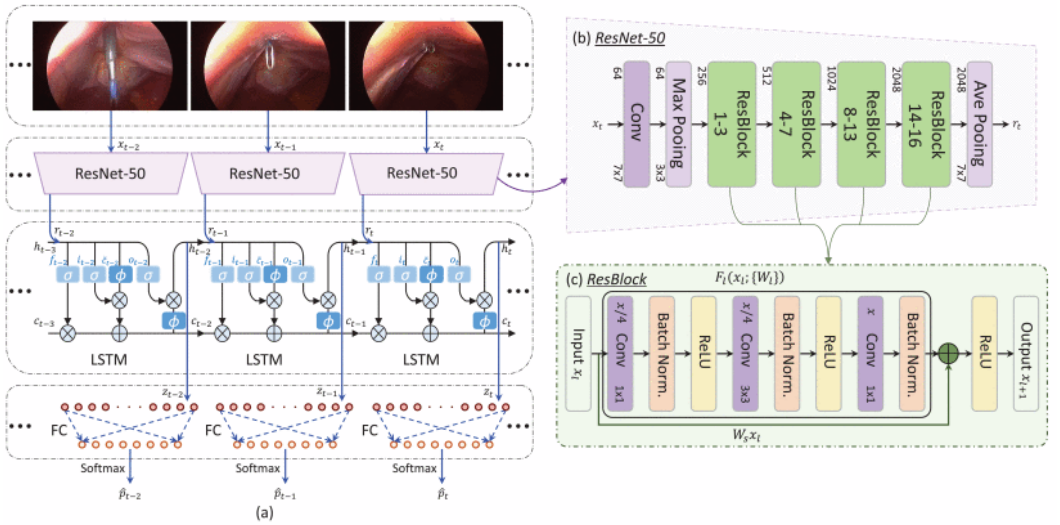
2.1 Surgical Activity Recognition

Surgical activity recognition is a cornerstone of SDS, aiming to automatically interpret and understand the content of surgical procedures from sensor data, most commonly video recordings. By providing a detailed, objective account of intraoperative events, surgical activity recognition enables a wide range of applications, including context-aware assistance in the operating room, objective surgical skill assessment, postoperative analysis, and automated documentation. The origins of surgical activity recognition date back to early work using statistical models like Hidden Markov Models (HMMs) to recognize high-level surgical phases from instrument usage data or handcrafted video features [11, 88]. These early methods modeled surgical procedures as a sequence of states, but often struggled to capture the complex, nuanced variations inherent in surgery [88]. With the advent of deep learning, the field has shifted towards data-driven approaches that can learn hierarchical representations directly from video, leading to significant performance gains [89, 90].

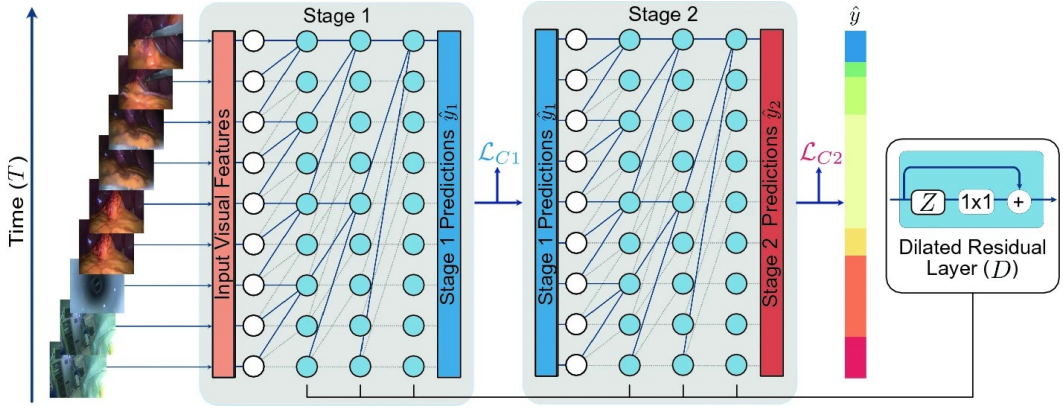
This section provides a review of the literature in surgical activity recognition, focusing on two main lines of research that differ in their level of granularity. Section 2.1.1 will discuss methods for recognizing coarse-grained surgical phases and steps, which represent high-level procedural stages. Section 2.1.2 will then delve into the more recent and challenging task of fine-grained action recognition and detection, which aims to identify detailed interactions between instruments, verbs, and tissues.

2.1.1 Coarse-grained Phase & Step Recognition

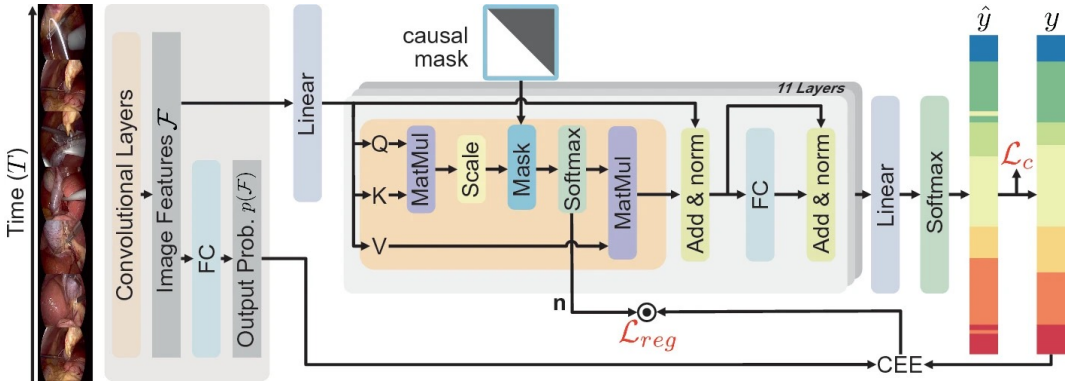
Surgical phase recognition involves segmenting a procedure into a sequence of distinct, chronologically ordered stages, such as "trocar insertion" or "gallbladder dissection" in a cholecystectomy. This task is fundamental for workflow analysis and intraoperative guid-



(a) LSTM-based method: SV-RCNet



(b) TCN-based method: TeCNO



(c) Transformer-based method: OperA

Figure 2.1: Different methods for phase and step recognition. (a) LSTM-based method: SV-RCNet. Image credit: [91]. (b) TCN-based method: TeCNO. Image credit: [64]. (c) Transformer-based method: OperA. Image credit: [92].

ance. The methodologies for phase recognition have evolved significantly, largely following the advancements in sequential data modeling in the broader computer vision community.

Early deep learning approaches replaced handcrafted features with learned representations from Convolutional Neural Networks (CNNs) and modeled the temporal dynamics using Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. For instance, Jin et al. [91] proposed SV-RCNet (Figure 2.1 (a)), which integrated a CNN and an LSTM to be trained end-to-end, effectively learning both spatial and temporal features from surgical videos for workflow recognition. While these models were effective, LSTMs can struggle to capture very long-term dependencies and their sequential nature makes them computationally intensive [93].

To address the limitations of RNNs, subsequent research turned to Temporal Convolutional Networks (TCNs). TCNs employ dilated convolutions to exponentially increase the receptive field, allowing them to capture long-range temporal relationships more efficiently and with greater parallelism than LSTMs [64, 94]. Czempiel et al. [64] were among the first to adapt a Multi-Stage TCN (MS-TCN) architecture for surgical phase recognition with their model, TeCNO (Figure 2.1 (b)), which demonstrated superior performance for online recognition. Their work showed that a two-stage refinement process improved results, though adding more stages could lead to overfitting [25]. Other studies have confirmed the advantages of TCN-based models over LSTM-based counterparts for both phase and step recognition tasks [25].

More recently, Transformer models, originally developed for natural language processing, have gained significant popularity for their ability to model global dependencies within a sequence [92]. The self-attention mechanism allows Transformers to weigh the importance of all frames in a video simultaneously, overcoming the sequential bottlenecks of RNNs and the fixed receptive fields of TCNs. Czempiel et al. [92] later proposed OperA (Figure 2.1 (c)), a

Transformer-based model that uses an attention regularization loss to encourage the model to focus on high-quality, informative frames, thereby improving prediction accuracy. Other works have explored different ways to leverage Transformers. For example, Ding et al. [95] presented a Segment-Attentive Hierarchical Consistency network (SAHC) that focuses on learning segment-level semantics to mitigate errors from ambiguous individual frames. To improve efficiency for long videos, some methods have proposed faster Transformer variants, such as SKiT by Liu et al. [96], which uses a key pooling operation to maintain global information with constant time complexity, making it suitable for real-time online recognition.

2.1.2 Fine-grained Action Recognition and Detection

While phase recognition provides a high-level overview of a procedure, a more detailed understanding is required for applications like automated skill assessment and detailed surgical process analysis. This has driven research towards fine-grained action recognition, which focuses on identifying specific, short-term actions or interactions.

A key step in this direction is surgeon action detection, which involves not only classifying an action but also localizing it within the video frame. To facilitate this research, Bawa et al. [97] introduced the SARAS Endoscopic Surgeon Action Detection (ESAD) dataset, a large-scale benchmark for detecting 21 distinct surgeon actions during prostatectomy procedures. The associated challenge provided baseline results using established object detection frameworks like Faster R-CNN [98] and YOLO [99], spurring the development of more specialized models [97].

To provide an even richer, more semantic description of surgical activities, Nwoye et al. [100] proposed representing actions as a triplet of <instrument, verb, target>, as shown in Figure 2.2. This formulation explicitly captures what tool is used, what action it performs,

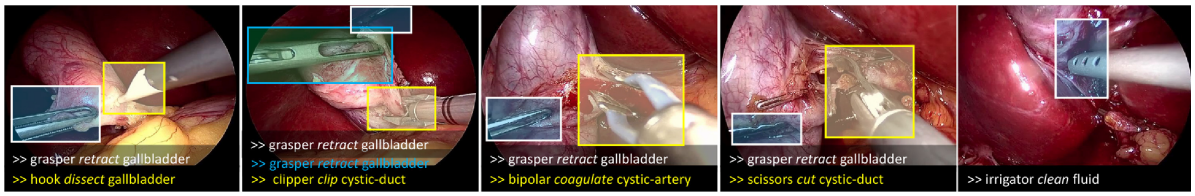


Figure 2.2: Few examples of action triplets in laparoscopic cholecystectomy surgical videos. Image credit: [44].

and what tissue it acts upon. Recognizing these triplets is challenging as it requires not only identifying each component but also correctly associating them [100]. To tackle this, Nwoye et al. [44] later developed the Rendezvous (RDV) model, which uses a combination of class activation-guided spatial attention and Transformer-inspired semantic attention to first detect the individual triplet elements and then model their relationships. Subsequent work has extended this by incorporating temporal information to produce smoother and more accurate triplet predictions over time [101].

Building on the concept of semantic triplets, recent work has aimed for even greater precision by including spatial localization. Lin et al. [102] introduced the task of instrument-tissue interaction quintuple detection, defining an interaction as a five-part structure: <instrument class, instrument bounding box, tissue class, tissue bounding box, action class>. They proposed QDNet, which uses a spatiotemporal attention layer and a graph-based prediction layer to reason about the relationships between localized instruments and tissues in cataract surgery videos [102]. This line of work was further advanced with their ITIDNet model, which better captures the relationships between elements within a single frame (intra-frame) and across consecutive frames (inter-frame) to improve detection accuracy in both cataract and cholecystectomy videos [103]. This progression from coarse phases to localized, fine-grained quintuples represents a clear trend towards developing a comprehensive and detailed automated understanding of surgical procedures.

2.1.3 Multi-task Learning in Surgical Scene

Multi-Task Learning (MTL) has emerged as a powerful paradigm in surgical scene analysis, as it can improve model generalization and performance by leveraging shared representations between related tasks [104]. The underlying principle is that training a single model to perform multiple, correlated objectives simultaneously encourages the learning of more robust and meaningful features than training separate models for each individual task.

A common application of MTL in surgery is the joint recognition of surgical phases and instrument presence, given their strong correlation within a well-defined surgical workflow. For instance, Jin et al. [105] proposed a Multi-Task Recurrent Convolutional Network (MTRCNet-CL) that utilized a shared encoder for visual feature extraction and two separate branches for phase recognition and tool detection. Crucially, they introduced a novel correlation loss to explicitly model the relationship between the two tasks at the prediction level, which was shown to significantly boost the performance of both [105]. This approach effectively demonstrated that enforcing high-level task correlation can lead to mutual performance benefits [105].

Building on this concept, other researchers have explored using MTL to analyze the surgical workflow at different levels of granularity. Ramesh et al. [25] developed a Multi-Task Multi-Stage Temporal Convolutional Network (MTMS-TCN) to jointly predict coarse-grained surgical phases and more fine-grained surgical steps. Their results indicated that the model benefited from the complementary nature of these tasks, outperforming single-task methods for both phase and step recognition [25]. This suggests that learning a hierarchical understanding of surgical procedures is an effective strategy.

The MTL paradigm has also been extended to incorporate a wider range of tasks for more holistic scene understanding. Some models [106] have been designed to simultaneously per-

form instrument segmentation and detect tool-tissue interactions, using a shared backbone to reduce computational costs compared to running two separate models [106]. Similarly, other works [107] have combined semantic segmentation of surgical tools with the detection of critical events, such as bleeding, within a single end-to-end framework [107]. These studies show the versatility of MTL in creating efficient, integrated systems capable of interpreting multiple complex aspects of the surgical scene in real-time.

2.2 Causal Inference

Deep learning models have achieved remarkable success by learning complex statistical correlations within large datasets. However, a fundamental limitation of this correlation-based learning is its susceptibility to data distribution shifts and spurious correlations [108]. A model may learn to associate a feature with a label not because of a true causal link [108], but because of a coincidental bias in the training data. For example, in surgical action recognition, a model might learn that the instrument "irrigation aspiration handpiece" is highly correlated with the high-frequency action "aspirate lens cortex". Due to this data bias, the model may fail to correctly recognize the rare action "inject saline", even when it is performed with the same instrument, because it has learned a spurious, non-causal link. Causal inference offers a mathematical framework to move beyond correlation and understand true cause-and-effect relationships [109], thereby improving model robustness and generalizability.

The foundational work in modern causal inference, largely developed by Judea Pearl [110], introduced the Structural Causal Model (SCM) as a way to represent causal assumptions about the world. An SCM uses a Directed Acyclic Graph (DAG) [108, 110], where nodes represent variables and directed edges represent direct causal relationships. This framework

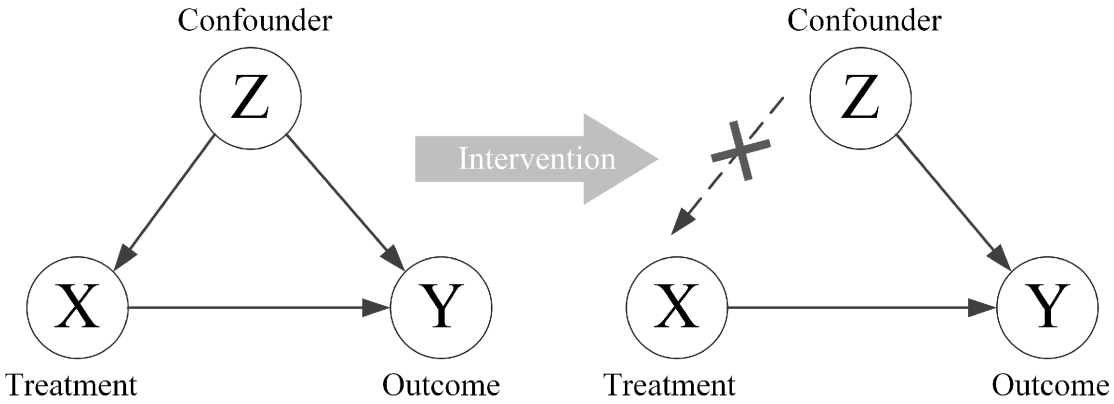


Figure 2.3: An example of causal graphs. Image credit: [113]. X: Treatment plays a direct causal role in determining the value of Outcome. Y: Outcome is the effect of the causal path. Z: Confounder is a common cause of both Treatment and Outcome.

provides a formal language to describe how variables are generated and interact. A key tool arising from this framework is the do-calculus, which allows for the estimation of the effect of an intervention, denoted as $P(Y|do(X = x))$ [109, 110]. This represents the distribution of an outcome Y if a variable X were forced to take a value x, which is distinct from the conditional probability $P(Y|X = x)$ that only reflects passive observation. By using tools like the backdoor and front-door adjustment criteria, researchers can identify and mitigate the effects of confounding variables-common causes that create spurious associations between other variables—to isolate the true causal effect of interest [111, 112]. Examples of causal graphs are illustrated in Figure 2.3.

In recent years, the computer vision community has increasingly adopted causal principles to address long-standing challenges where models exploit dataset biases instead of learning the intended concepts. These spurious correlations often stem from confounding variables, which are factors that influence both the input features and the output labels, leading to poor generalization.

A prominent application area is long-tailed classification, where models tend to be biased towards head classes with abundant samples. Tang et al. [114] identified the Stochastic

Gradient Descent (SGD) momentum as a key confounder that harms tail-class performance. They proposed a causal framework to disentangle the harmful and beneficial effects of momentum, using causal intervention during training to remove the confounding bias while retaining its positive impact on representation learning. More recently, as Vision Transformers (ViTs) have become prevalent, researchers have adapted causal models to address their unique properties. For instance, TSCNet was proposed to discover fine-grained causal associations through multi-scale interventions, addressing the challenge that ViT's global feature representation can obscure the signals needed to classify visually similar tail classes [115].

Weakly-Supervised Semantic Segmentation (WSSS) is another area where causal inference has proven effective. A primary challenge in WSSS is that models often use co-occurring context as a shortcut for classification, resulting in inaccurate object localization maps. Zhang et al. [116] introduced a causal framework that models the confounding effect of context. By applying causal intervention (specifically, the backdoor adjustment), their method, CONTA, can remove the confounding bias from the image representation, leading to the generation of higher-quality pseudo-masks for training the segmentation model.

In multi-modal tasks like Visual Question Answering (VQA), models often suffer from language bias, where they learn to answer questions based on question-answer priors in the training data while ignoring the visual content. To combat this, several methods have employed counterfactual reasoning. Chen et al. [117] proposed a model-agnostic training scheme that synthesizes counterfactual samples by masking critical objects in the image or words in the question, forcing the model to be more visually-grounded and question-sensitive. Wen et al. [118] also used a counterfactual inference framework to reduce language bias. Similarly, in video-based QA, where unobserved confounders can create spurious correlations between video content and answers, Li et al. [119] developed a framework using front-door adjustment to remove the confounding effect without needing to observe the confounder

itself.

The application of causal inference extends to other visual tasks as well. In video moment retrieval, Yang et al. [120] proposed a causality-inspired framework to address temporal location biases, improving query-video matching through causal intervention. For zero-shot action recognition, Wang et al. presented a method using causal inference to generate features for unseen actions, effectively handling confounders [112]. These works collectively demonstrate that by explicitly modeling and intervening on the causal mechanisms of data generation, it is possible to build more robust and generalizable vision systems.

2.2.1 Causal Inference in Medical Scene

The principles of causal inference are particularly vital in the medical domain, where model reliability, fairness, and interpretability are paramount for clinical translation [121]. Medical AI models are susceptible to biases arising from data collection, patient demographics, and the inherent co-occurrence of anatomical structures and pathologies, which can lead to erroneous diagnostic conclusions [121].

In medical image segmentation, a significant challenge is the severe co-occurrence of different organs or tissues in training data, which can confuse weakly-supervised models. For example, a model might struggle to distinguish the liver from the right kidney because they frequently appear together. To address this, Chen et al. [122] proposed Causal CAM (C-CAM), a method for weakly supervised medical image segmentation. C-CAM is motivated by two causal chains: one linking image content to its category and another linking anatomical structure to organ shape. By integrating anatomical constraints through a causal lens, the method mitigates the co-occurrence problem and generates more accurate pseudo-masks for segmentation [122].

For medical diagnosis, causal frameworks are being developed to ensure that predictions are based on genuine pathological features rather than being confounded by sensitive attributes like patient race or gender [123]. It has been shown that models can unintentionally learn to use demographic information inferred from image patterns as a shortcut for diagnosis, raising serious fairness concerns [123]. To mitigate this, causal modeling frameworks have been proposed to control the influence of such attributes on predictions, ensuring that decisions are primarily informed by clinically relevant information [123].

Causal inference also enhances explainability in medical AI. Counterfactual explanations, which show how an input image would need to change to alter a model's prediction, are being explored to improve clinicians' trust and understanding [124]. For example, a counterfactual might show what features in a chest X-ray would need to be different for a diagnosis to change from "pneumonia" to "normal". Studies have shown that such explanations can help users critically review AI suggestions and reduce over-reliance on incorrect AI outputs, fostering more effective human-AI collaboration in clinical decision-making [125]. Furthermore, in the context of medical report generation, which suffers from similar multi-modal biases as VQA, causal intervention has been used to disentangle spurious correlations between visual features and linguistic descriptions, leading to more accurate and reliable reports [11, 126]. The systematic application of causal reasoning holds great promise for developing medical AI systems that are not only accurate but also fair, robust, and trustworthy for real-world clinical use.

2.3 Foundation Model in Medical Scene

The paradigm of artificial intelligence has been reshaped by the emergence of Foundation Models (FMs), which are large-scale models pre-trained on vast, diverse datasets [127, 128].

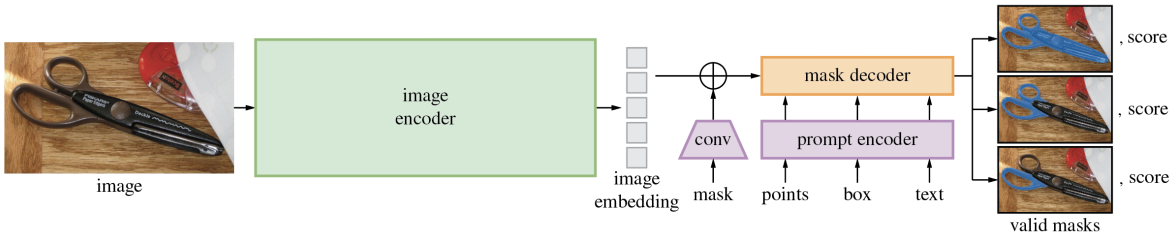


Figure 2.4: Architecture of the Segment Anything Model (SAM). Image credit: [131].

Unlike traditional models designed for a single, specific task, foundation models learn general-purpose representations primarily through self-supervised learning, though this pre-training can also incorporate other paradigms such as large-scale supervised training, human-in-the-loop feedback, or reinforcement learning. These models can then be adapted to a wide range of downstream applications with minimal task-specific training [129, 130]. This adaptability, often realized through zero-shot or few-shot learning via prompting, offers significant advantages, including accelerated development time, reduced need for extensive labeled datasets, and enhanced performance on novel tasks [129, 130]. The success of foundation models in natural vision has directly inspired their exploration within the medical domain [129]. These general-purpose models have set remarkable performance benchmarks; for example, the Segment Anything Model (SAM) was trained on a dataset of over one billion masks to perform promptable, class-agnostic segmentation in a zero-shot manner [128, 131] (presented in Figure 2.4). Concurrently, multimodal foundation models like CLIP have enabled novel forms of reasoning by learning a shared embedding space for images and text [132, 133]. These models exemplify the power of large-scale pre-training and form the conceptual basis for the medical foundation models discussed next.

However, the direct application of these models to medical imaging is challenging due to the significant domain shift; medical images possess unique characteristics across different modalities (e.g., CT, MRI, endoscopy) and often feature subtle, low-contrast pathologies not present in natural images [134, 135]. Initial studies evaluating the zero-shot performance

of SAM on various medical segmentation tasks reported promising but inconsistent results, indicating that direct application is often suboptimal compared to specialist models trained on in-domain data [136, 137]. A review by Fan et al. [134] highlighted that while SAM shows potential, modifications are necessary to handle the intricacies of medical data.

Consequently, recent research has focused on adapting foundation models for the medical scene. One successful strategy involves fine-tuning pre-trained models on curated medical datasets. For instance, several works have proposed adapting SAM by fine-tuning its mask decoder or inserting lightweight, learnable adapter modules to improve segmentation accuracy on specific anatomical structures or lesions [135, 138]. This approach leverages the powerful pre-trained encoder while tailoring the model to the specific features of medical images, a method that has proven effective for both 2D and 3D medical image segmentation [135, 138]. These adaptations have led to the development of specialized models that significantly outperform the original general-purpose foundation models on medical benchmarks [138].

Within the clinical context, the surgical scene presents a particularly compelling use case for foundation models, especially for automated workflow analysis from endoscopic videos [49]. Recent studies have begun to integrate FMs to enhance surgical scene understanding, including tasks like instrument tracking, anatomical segmentation, and surgical phase recognition [49]. For example, some approaches [81] leverage foundation models to provide robust visual features that are then used by temporal models to recognize the distinct phases of a surgical procedure [81]. Chen et al. [139] proposed a text-driven adaptation method that allows a foundation model to perform few-shot surgical workflow analysis by aligning visual features with textual descriptions of surgical actions and phases. This mitigates the need for large-scale video annotations and demonstrates the potential of multimodal foundation models to create context-aware systems that can improve surgical safety and efficiency [49].

2.4 Thesis positioning

While the review in Section 2.1, 2.2, and 2.3 highlights significant progress, this thesis identifies and addresses several key limitations in the current literature to advance multi-level surgical scene understanding. We position our contributions in the following four areas:

The review in Section 2.1.2 highlights a growing interest in fine-grained action recognition [97, 102, 140]. However, a majority of work remains focused on coarse-grained phase recognition [22, 61], and existing fine-grained detection models often yield modest classification accuracy [39]. Furthermore, a critical deficiency noted in the literature is the lack of robust prediction uncertainty analysis [62, 63], which is essential for clinical adoption. This thesis addresses these gaps in Chapter 4 by proposing the Anchor-context Action Detection Network (ACTNet). This framework introduces a Class Conditional Diffusion (CCD) module specifically to refine action predictions and, for the first time, to provide reliable confidence scores, a key requirement for clinical trust.

Chapter 5 builds upon this by tackling the lack of a holistic, multi-level understanding. The review in Section 2.1.1 and 2.1.3 shows that most frameworks target isolated tasks [20, 64] or simple multi-task combinations [25, 105], failing to exploit the inherent hierarchical relationships between phases, steps, and actions [21, 65]. To address this, this thesis proposes a Multi-level Semantic Scene Understanding (MSSU) framework, the Hierarchical Context Transformer (HCT). HCT explicitly models this hierarchy through a novel Hierarchical Relation Aggregation Module (HRAM) and enhances representation learning via Inter-task Contrastive Learning (ICL), demonstrating a more comprehensive understanding than the SOTA methods.

Chapter 6 addresses a more fundamental limitation: the tendency of models to learn spurious correlations from visual data, as discussed in Section 2.2 [66, 67]. This reliance

on statistical shortcuts rather than true causal links negatively impacts model robustness and generalization [86]. While causal inference is an active field [109, 114, 122], its application to deconfound multi-level surgical video representations is underexplored. This thesis introduces a causality-inspired framework, Deconfounded Surgical Video Representation Learning (DSVRL). By incorporating a Subject-Centric Module (SCM) and an Activity Causal Intervention Module (ACIM) based on backdoor adjustment, this framework mitigges confounding biases and learns the true causal effects, a significant step beyond the correlation-based learning of conventional models.

Finally, Chapter 7 tackles the critical challenge of model generalization across different clinical environments, a phenomenon known as domain shift [69]. As noted, models often fail to transfer between different hospitals or surgeons [69]. While foundation models (Section 2.3) show promise for generalization [127, 128, 131], their application to standardize multi-center surgical workflow analysis is a key gap. This thesis develops the Comprehensive Ophthalmology Surgical (COS) Model, a foundation model for multi-level workflow recognition. By leveraging hierarchical spatiotemporal feature extraction (ViT) and validating extensively on multi-center datasets, this work directly addresses the need for robust, scalable, and generalizable workflow analysis tools, moving beyond the single-center limitations of many previous studies.

Chapter Three

Multi-level Surgical Video Dataset Construction

Deep learning models fundamentally operate as function approximators, learning from historical or available observations. The volume and quality of these observations are paramount, directly influencing the ability of these models to accurately estimate the underlying function, particularly for complex tasks. Consequently, models with millions of parameters, common in fields like computer vision, necessitate large, meticulously labeled datasets. These datasets enable the model parameters to effectively capture salient representations of the domain, thereby achieving optimal performance on the intended task [141–143]. The critical role of dataset size and representativeness in model accuracy and computational efficiency, especially within resource-sensitive fields like medical imaging, continues to be an active area of investigation [141]. Indeed, the availability of high-quality, large-scale datasets is often a determining factor in the success of deep learning applications [144].

In the broader computer vision community, the introduction of extensive datasets has consistently catalyzed progress across a multitude of visual tasks. Seminal datasets such as ImageNet [145], COCO [146], and Kinetics [147], among others, have been instrumental in advancing areas like image classification, object detection, semantic segmentation, and action recognition. These resources have not only facilitated the development of more sophisticated algorithms but have also provided robust benchmarks for comparing model performance, thereby fostering innovation and pushing the boundaries of what is achievable in visual understanding [143]. The impact of these foundational datasets is evident in the rapid evolution of deep learning architectures and their successful application in diverse real-world scenarios [143].

Similarly, the field of surgical vision has recognized the indispensable need for high-quality datasets to drive advancements in computer-assisted interventions and surgical data science [78, 149]. Efforts have led to the creation of several public datasets targeting various surgical procedures and analytical tasks. For instance, Cholec80 has been widely used for laparoscopic cholecystectomy phase recognition and, more recently, for training AI models with high

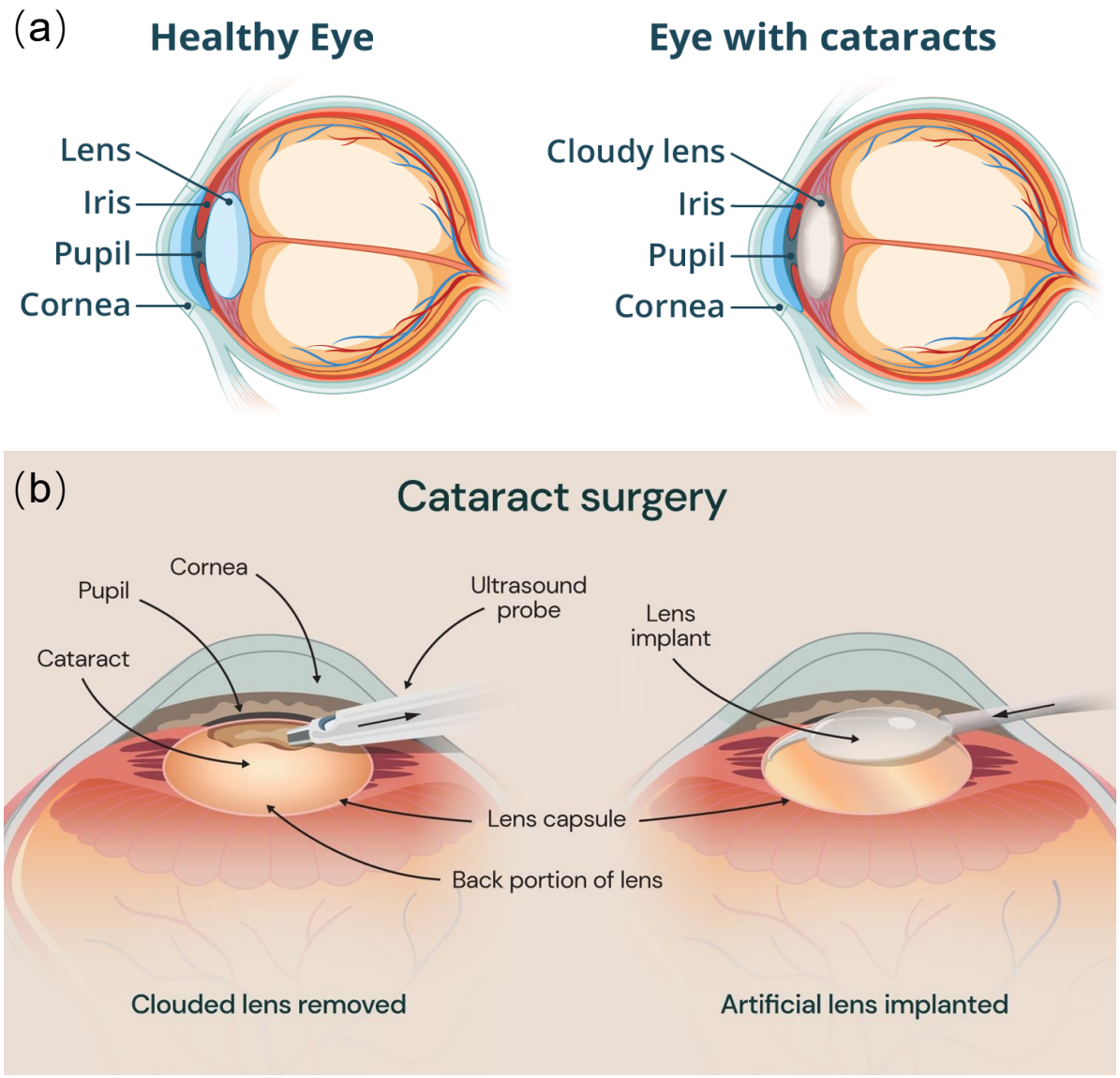


Figure 3.1: Cataract eye and cataract surgery illustration. (a) Comparison between a healthy eye and a cataract eye, highlighting the difference in lens status. Image credits: Healthdirect. (b) Diagram of cataract surgery, depicting the removal of the clouded lens and implantation of an artificial lens. Image credits: Guy's and St Thomas' Specialist Care [148].

accuracy [40, 150]. The CATARACTS dataset has been pivotal for developing algorithms for phase discrimination and instrument tracking in cataract surgeries [32, 41]. Other notable datasets include HeiSurF [151], focusing on phase recognition, and the EndoVis challenge datasets, which often feature tasks like instrument segmentation and tracking [152, 153]. While these datasets, along with others like CaDIS for cataract surgery image segmentation [154, 155] and those for action recognition [156], have spurred significant research, many focus on specific levels of activity or a limited range of procedures [47]. The annotation of surgical video itself presents unique challenges, requiring expert knowledge and standardized approaches to ensure data quality and inter-rater reliability, which are crucial for training robust AI models [157]. There remains a pressing need for more comprehensive datasets that capture the multifaceted nature of surgical workflows [154, 158].

To this end, in this chapter, we present a multi-level, multi-granularity cataract surgery dataset that includes annotations for three types of activities: phases, steps, and fine-grained actions. We also briefly introduce a multi-granularity public endoscopic dataset.

3.1 Cataract Dataset

3.1.1 Ethical approval

The cataract dataset utilized in this study, comprising cataract surgical videos from Shenzhen First People's Hospital, was collected under institutional review board approval (reference number: LL-KY-2020220) with appropriate informed consent procedures in accordance with the Declaration of Helsinki. Comprehensive de-identification protocols were implemented prior to any computational processing or model development activities to ensure patient privacy and data security compliance. All personally identifiable information, including but

not limited to patient demographics, case identifiers, and timestamp metadata, was removed following institutional standards for anonymization of surgical video content. The complete de-identification workflow was independently verified by the institutional data protection officer at the contributing center to confirm adherence to both local regulatory requirements and international data protection standards.

3.1.2 Cataract Surgery

A cataract is a prevalent ocular condition characterized by the opacification of the eye's natural lens, leading to a progressive decline in visual acuity [148]. Understanding the anatomical differences between a healthy eye and an eye afflicted by cataracts, as well as the intricacies of surgical correction, is crucial.

Figure 3.1 illustrates a healthy eye wherein the lens, situated behind the iris and pupil, is transparent, allowing light to pass unimpeded to the retina. Key anatomical structures include the cornea, the eye's clear outer layer that refracts light; the iris, the colored part that controls the size of the pupil; and the pupil, the central opening that regulates light entry. In contrast, an eye with cataracts, also depicted in Figure 3.1 (a), features a cloudy or opaque lens. This cloudiness scatters and blocks light, resulting in blurred vision and reduced visual clarity, a hallmark of cataract pathology [148].

Cataract surgery is the primary treatment to restore vision. Figure 3.1 (b) provides a simplified overview, showing the clouded lens being removed and replaced by an artificial lens implant. A more detailed procedural sequence, commonly involving phacoemulsification, is illustrated in Figure 3.2. Phacoemulsification is a sophisticated technique that allows for cataract removal through a small incision, promoting faster recovery and minimizing astigmatism; it remains the gold standard for cataract surgery [159].

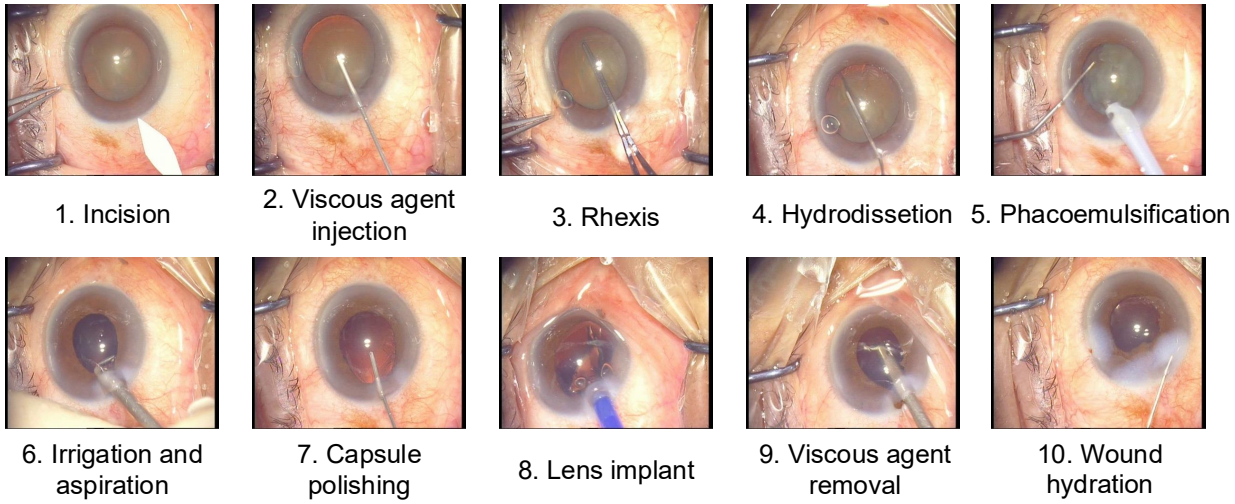


Figure 3.2: Surgical procedure demonstration based on real surgical data. Data source: Our cataract dataset.

The surgical procedure, as demonstrated in Figure 3.2, is a multi-step process. It commences with a precise micro-incision, typically in the cornea, followed by the injection of a viscous agent to maintain anterior chamber depth and protect delicate intraocular tissues [160]. A critical circular opening, known as a capsulorhexis, is then created in the anterior capsule of the lens. This allows for hydrodissection, where fluid is injected to separate the lens nucleus from the capsule, facilitating its removal. The core of the surgery involves phacoemulsification, where an ultrasonic probe emulsifies the hardened, cloudy lens nucleus into smaller fragments, which are simultaneously aspirated from the eye [161]. Following this, any remaining softer lens cortex is meticulously removed using irrigation and aspiration, and the inner surface of the lens capsule may be polished. An artificial Intraocular Lens (IOL) is then folded and inserted into the empty capsular bag, replacing the natural lens. The procedure concludes with the thorough removal of the viscous agent and hydration of the corneal wound to ensure a self-sealing closure [162]. This comprehensive surgical approach aims to replace the opacified natural lens with a clear artificial implant, thereby restoring transparency to the optical pathway and significantly improving the patient's visual function.

3.1.3 Label Definition

To facilitate a comprehensive understanding of cataract surgery for computational analysis, we have developed a hierarchical annotation scheme. This annotation framework was established in collaboration with two senior ophthalmic experts. The data annotation process was then conducted as follows: first, both experts annotated the surgical videos independently; subsequently, they held discussion sessions to review and resolve any discrepancies between their annotations, thereby ensuring a final, consensus-based ground truth.

Surgical procedures were segmented into three levels of detail: five distinct phases, eleven procedural steps, and fifty specific actions, progressing from a coarse overview to fine-grained operational details. The definitions for these labels are presented in Section A.1. Based on these definitions, experienced ophthalmologists performed frame-by-frame annotation of our dataset, with a secondary review by another ophthalmologist ensuring accuracy and consistency. This multi-level labeling approach provides a comprehensive description of the surgical scenario at different semantic depths, which can better assist surgeons in grasping the full context of the operation, thereby aiding decision-making and enhancing surgical efficiency.

3.1.4 Dataset Statistics

Our cataract dataset, central to this study, comprises 61 cataract surgery videos (85,659 frames at 1 fps, 720 $\times$ 576 resolution). Ophthalmologists annotated each frame for 5 surgical phases and 11 steps. For experimentation, the data is split at the patient level into training (35 videos), validation (6 videos), and test sets (20 videos), ensuring that videos from the same patient do not appear in different sets. A subset of 20 videos (23,580 frames, 28,428 action instances) is further annotated for action detection, detailing 49 action and

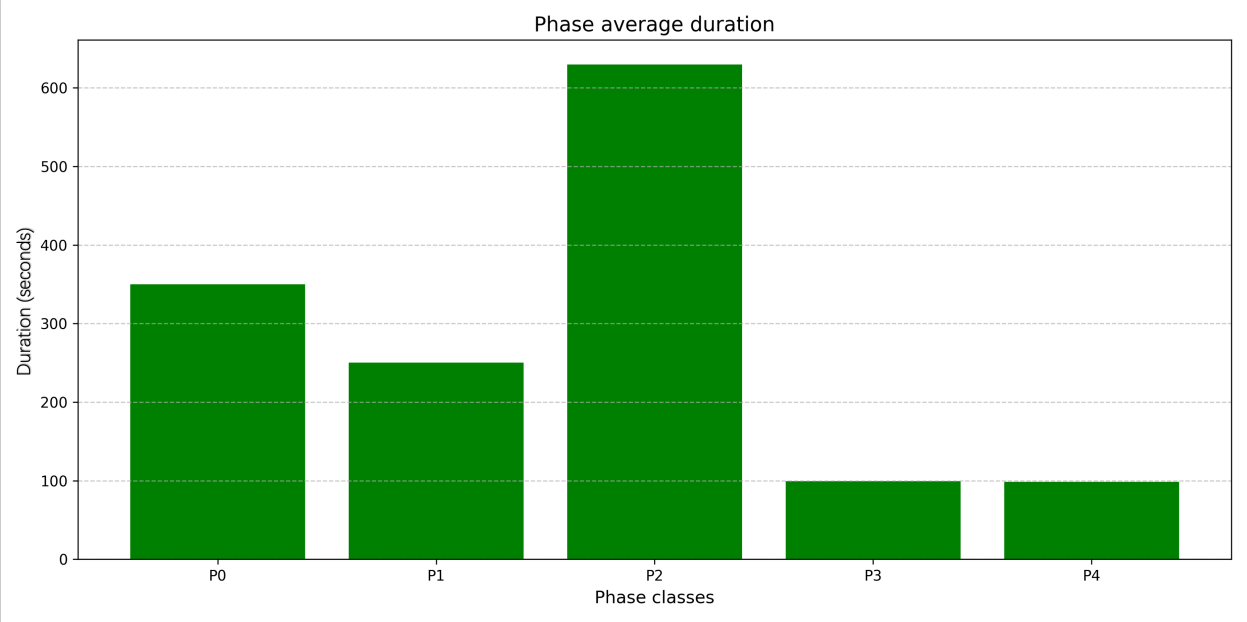


Figure 3.3: Average duration of phases in the Cataract dataset.

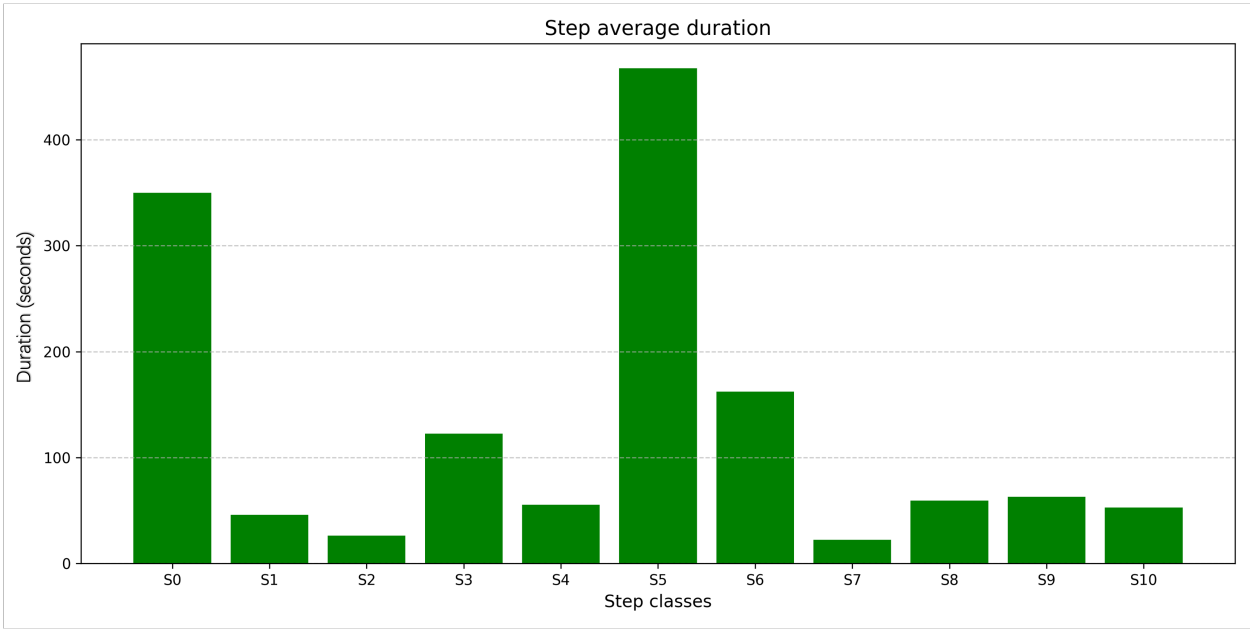


Figure 3.4: Average duration of steps in the Cataract dataset.

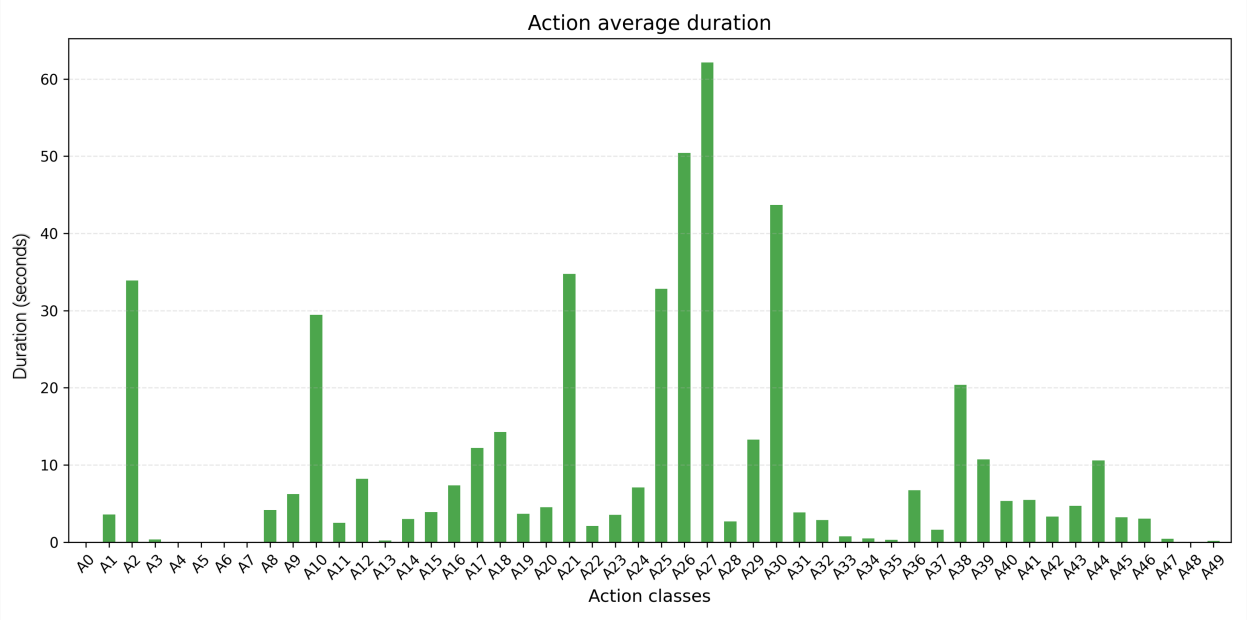


Figure 3.5: Average duration of actions in the Cataract dataset.

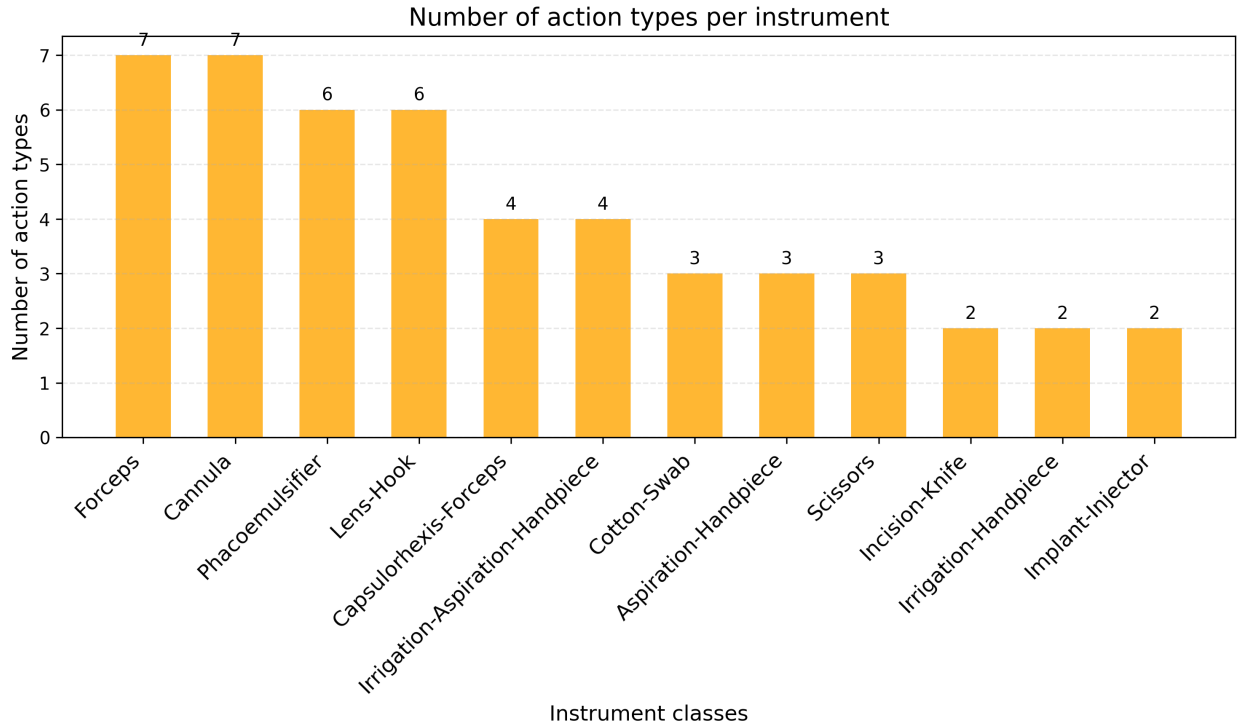


Figure 3.6: Number of action types performed by different surgical instruments in the Cataract dataset.

13 instrument labels with bounding boxes, which is similarly split at the patient level into training (12 videos), validation (3 videos), and test sets (5 videos).

A key challenge presented by this dataset is significant class imbalance across all annotation granularities, mirroring real-world surgical data. Figure 3.3 reveals substantial disparities in the average durations of surgical phases. Notably, Phase P2 (e.g., phacoemulsification) is considerably longer (over 600 seconds in aggregated frame time) than others like P3 and P4 (around 100 seconds), potentially biasing model learning towards dominant phases. This imbalance is exacerbated at the surgical step level (Figure 3.4). Steps S0 and S5 show the longest average durations (over 300 and 450 seconds respectively), while many other steps (e.g., S2, S7) are markedly shorter (under 50 seconds). Such variations pose difficulties for models to learn underrepresented steps effectively.

The issue of class imbalance is most acute for action detection (Figure 3.5). Many of the 49 action classes exhibit very brief average durations (often less than 10 seconds of aggregated frame time), while a few, such as A28, are significantly longer (exceeding 60 seconds). This severe skew necessitates robust strategies to detect both frequent and infrequent actions. Furthermore, the complexity of action detection is amplified by surgical practice. As indicated in Figure 3.6, instruments like Forceps and Cannula are associated with up to 7 action types, highlighting that single instruments perform multiple functions. Critically, multiple actions often occur simultaneously, involving one or more instruments. This co-occurrence, coupled with the multi-functional nature of instruments, makes precise spatio-temporal localization and classification of individual actions highly demanding.

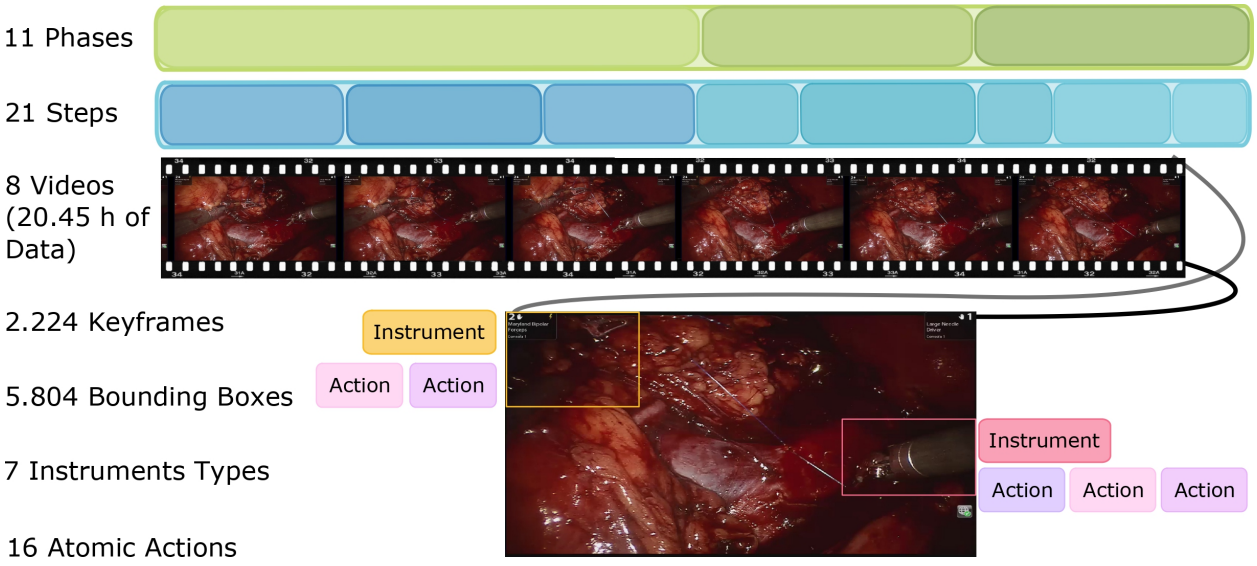


Figure 3.7: Illustration of PSI-AVA dataset. Image credit: [21].

3.2 Endoscopic Dataset

Validating our proposed method on public datasets is an essential way to demonstrate its effectiveness. Therefore, this section introduces the primary multi-granularity public endoscopic dataset used in our work, PSI-AVA [21]. Endoscopic surgery, such as the robot-assisted radical prostatectomy procedures captured in the PSI-AVA dataset, is performed inside a body cavity using an endoscope and specialized instruments. This approach differs significantly from microscopic surgery, such as cataract operations. Microscopic surgery typically involves a highly magnified but smaller, more constrained field of view focused on a specific anatomical site. In contrast, endoscopic surgery presents a wider and more comprehensive view of the surgical field, capturing a complex environment with dynamic, coordinated interactions between multiple instruments and various tissues. The da Vinci system used for the surgeries in PSI-AVA, for example, provides a magnified, high-definition 3D view of the entire surgical area. The distinct characteristics of these surgical video types mean their dataset features are markedly different. By testing and validating our method on two fundamentally different types of surgical datasets, we can more convincingly demonstrate its

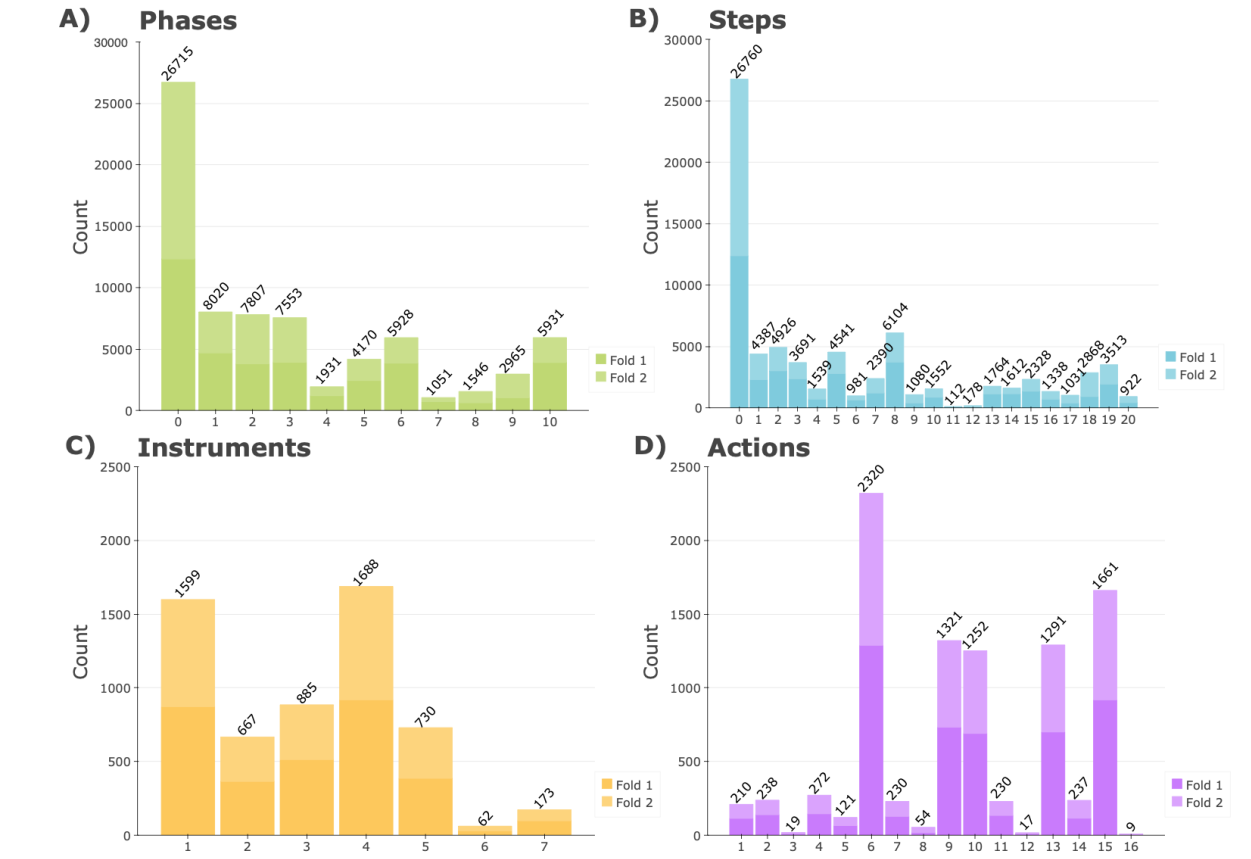


Figure 3.8: PSI-AVA Statistics. Number of annotations for each class of the recognition and detection tasks. Colors denote the distribution in the fold partition. Image credit: [21].

robustness and high adaptability.

PSI-AVA dataset is a benchmark specifically designed to promote a holistic understanding of surgical scenes. It was created from eight robot-assisted radical prostatectomy surgeries performed by three expert surgeons, totaling over 20 hours of video. The dataset's core strength lies in its multi-level annotations, which deconstruct the complex surgical procedure into both long-term and short-term activities, as shown in Figure 3.7.

This hierarchical annotation framework enables a comprehensive analysis of the surgical workflow: 1) Long-Term Reasoning: This level captures the overall progress of the surgery. It includes annotations for 11 distinct surgical Phases and 20 more granular Steps that

constitute those phases. These are annotated densely for every second of video, providing a continuous timeline of the procedure. 2) Short-Term Reasoning: This level focuses on the fine-grained actions occurring at specific moments. It provides spatial and categorical information for instruments and their activities. This includes bounding box annotations for 7 types of surgical Instruments and, notably, a novel Atomic Action Recognition task with 16 action classes. These atomic actions are designed to be general and fundamental, describing what an instrument is doing (e.g., "cutting", "grasping") independent of the specific surgical context. In total, the dataset contains 73,618 keyframes annotated for phases and steps. For the more detailed short-term tasks, 2,238 keyframes are annotated with 5,804 instrument instances, each with a category label and associated atomic actions. This rich, multi-layered structure allows for studying the intrinsic relationships between high-level surgical goals and the low-level instrument dynamics.

An analysis of the instance distribution within the PSI-AVA dataset, as depicted in Figure 3.8, reveals several intrinsic challenges inherent to modeling surgical workflow. A prominent characteristic across all four annotation granularities—phases, steps, instruments, and actions—is the severe data imbalance, commonly known as the long-tail problem. For instance, in both the phase and step recognition tasks, a single category (labeled "0"), likely corresponding to an "Idle" or preparatory state, dominates the dataset with over 26,000 instances. In contrast, many other clinically significant categories are substantially less frequent, with some step and action classes having fewer than 100 instances. This significant skew poses a considerable challenge for training deep learning models, as it can lead to a strong bias towards the majority classes, potentially hindering the accurate recognition of rare but critical surgical events.

Furthermore, the dataset's complexity is amplified by the sheer number of categories at the finer levels of granularity. With 21 distinct steps and 16 atomic actions, the model must learn a high-dimensional and nuanced feature space to differentiate between subtle variations

in surgical execution. This suggests that the surgical procedures involve frequent and rapid switching between a wide variety of specific maneuvers. Compounding this challenge is the concurrency of events; the underlying annotation protocol allows for multiple atomic actions to be assigned to a single instrument simultaneously. Consequently, these factors—the pronounced class imbalance, the large number of categories at multiple granularities, and the potential for concurrent actions—establish PSI-AVA as a demanding benchmark. Overcoming these challenges is essential for developing robust models capable of achieving a truly holistic understanding of the surgical scene.

Chapter Four

Fine-grained Surgical Video Scene Understanding

This chapter is based on the following publication:

ACT-Net: Anchor-context Action Detection in Surgery Videos [163], which was published in the International Conference on Medical Image Computing and Computer-Assisted Intervention (2023).

Chapter 3 established the foundational dataset for this chapter, highlighting the significant challenges in surgical video, such as severe class imbalance and the complexity of multi-level annotations. A key challenge identified is the difficulty in robustly recognizing fine-grained actions, which are often brief, concurrent, and visually subtle. Standard detection models struggle with this complexity. Therefore, this chapter addresses this specific gap by proposing a novel methodology focused explicitly on improving fine-grained surgical action detection, which serves as the first crucial step towards building a holistic, multi-level understanding framework.

4.1 Introduction

Surgery is often an effective therapy that can alleviate disabilities and reduce the risk of death from common conditions [164]. While surgical procedures are intended to save lives, errors within the surgery may bring great risks to the patient and even cause sequelae [165], which emphasizes the development of a computer-assisted system. A context-aware assistant system for surgery can not only decrease intraoperative adverse events, and enhance the quality of interventional healthcare [12], but also contribute to surgeon training, and assist procedure planning and retrospective analysis[166].

Designing intelligent assistance systems for operating rooms requires an understanding of surgical scenes and procedures [167]. Most current works pay attention to phase and step recognition [40, 168], which is to get the major types of events that occurred during the surgery. They merely provided very coarse descriptions of scenes. As the granularity of action increases, the clinical utility becomes more valuable in providing an accurate depiction of detailed motion [44, 102]. Some studies focus on fine-grained action recognition by modelling action as a group of the instrument, its role, and its target anatomy and capturing their

associations [106, 169]. Recognizing targets in different methods is dependent on different surgical scenarios and it also significantly increases the complexity and time consumption for anatomy annotation [170]. In addition, although most existing methods can provide accurate action positions, the predicted action class is often inaccurate. Moreover, they do not provide any information about the reliability of their output, which is a key requirement for integrating into assistance systems of surgery [171]. Thus, we propose a reliable surgical action detection method in this chapter, with high-accuracy action predictions and their confidence.

Mistrust is a major barrier to deep-learning-based predictions applied to clinical implementation [172]. Existing works measuring the model uncertainty [173, 174] often need several-time re-evaluations, and store multiple sets of weights. It is hard for them to apply to surgery assistance applications to get confidence for each prediction directly[175], and they are limited to improving prediction performance. Conditional diffusion-based generative models have received significant attention due to their ability to accurately recover the full distribution of data guided by conditions from the perspective of diffusion probabilistic models [176]. However, they focus on generating high-resolution photo-realistic images. Instead, after observing our surgical video dataset, our conditional diffusion model aims to accurately reconstruct class distribution. We also access the estimation of confidence with the stochastic nature of the diffusion model.

Here, to predict accurately micro-action (fine-grained action) categories happening every moment, we achieve it with two modules. Specifically, a novel anchor-context module for action detection is proposed to highlight the spatio-temporal regions that are interacted with the anchors (we extract instrument features as anchors), which includes surrounding tissues and movement information. Then, with the constraints of class distributions and the surgical videos, we propose a conditional diffusion model to cover the whole distribution of our data and to accurately reconstruct new predictions based on full learning. Furthermore,

our class conditional diffusion model also accesses uncertainty for each prediction, through the stochasticity of outputs.

We summarize our main contributions as follows: 1) We develop an Anchor-context Action Detection Network (ACTNet), including an Anchor-Context Detection (ACD) module and a Class Conditional Diffusion (CCD) module, which combines three tasks: i) where actions locate; ii) what actions are; iii) how confident our model is about predictions. 2) For ACD module, we develop a Spatio-Temporal Anchor interaction Block (STAB) to spatially and temporally highlight the context related to the extracted anchor, which provides micro-action location and initial class. 3) By conditioning on the full distribution of action classes in the surgical videos, our proposed CCD model reconstructs better class prototypes in a stochastic fashion, to provide a more accurate estimations and push the assessment of the model confidence in its predictions. 4) We carry out comparison and ablation study experiments to demonstrate the effectiveness of our proposed algorithm based on cataract surgery.

4.2 Methodology

The overall framework of our proposed ACTNet for reliable action detection is illustrated in Figure 4.1. Based on a video frame sequence, the ACD module extracts anchor features and aggregates the spatio-temporal interactions with anchor features by the proposed STAB, which generates action locations and initial action class distributions. Then considering the full distribution of action classes in surgical videos, we use the CCD module to refine the action class predictions and get confidence estimations.

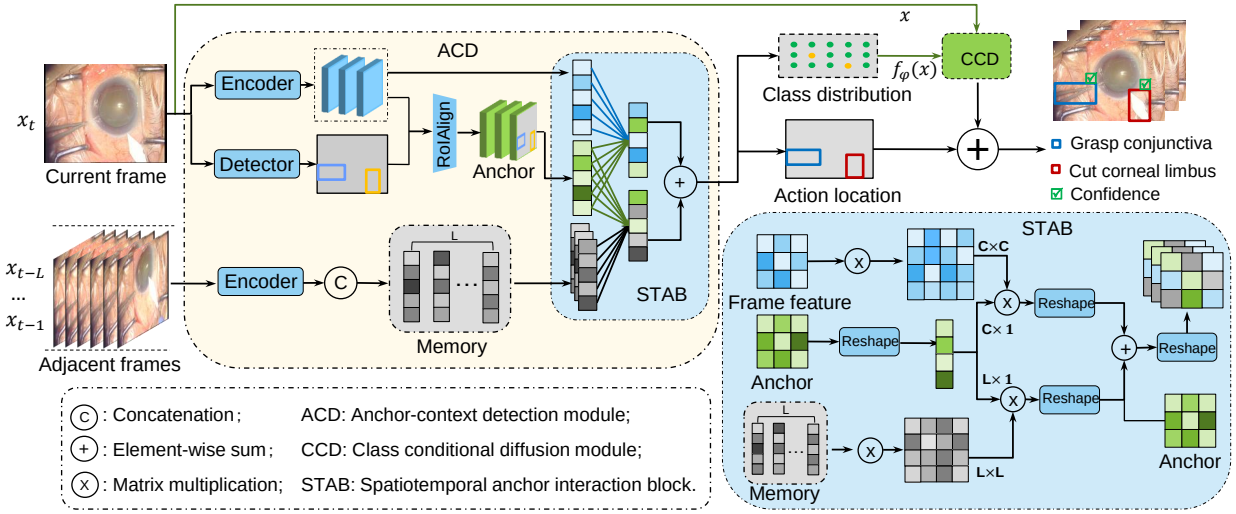


Figure 4.1: The pipeline of our Anchor-context Action Detection Network (ACTNet) includes Anchor-Context Detection (ACD) and Class Conditional Diffusion (CCD) modules.

4.2.1 Our Anchor-Context Detection Module

4.2.1.1 Anchor Extraction

Assuming a video X with T frames, denoted as $X = \{x_t\}_{t=1}^{T}$, where x_t is the t-th frame of the video. The task of this work is to estimate all potential locations and classes $P = \{box_n, c_n\}_{n=1}^{N}$ for action instances contained in video X, where box_n is the position of the n-th action happened, c_n is the action class of n-th action, and N is the number of action instances. For video representation, this work tries to encode the original videos into features based on the backbone ResNet50 [177] network to get each frame's feature $F = \{f_t\}_{t=1}^{T}$.

In surgical videos, the instruments, as action subjects, are significant to recognize the action. For instrument detection, it is very important but not very complicated. Existing excellent object detection method like Faster R-CNN [98] is enough to obtain results with high accuracy. After getting the detected instrument anchors, RoIAlign is applied to extract the instrument features from frame features. The instrument features are denoted as $I =$

$\{i_t\}_{t=1}^{T}$. Since multiple instruments exist in surgeries, our action detection needs to solve the problem that related or disparate concurrent actions often lead to wrong predictions. Thus, in this chapter, we propose to provide action location and class considering the spatio-temporal anchor interactions in the surgical videos, based on STAB.

4.2.1.2 Spatio-Temporal Action Interaction Block (STAB)

A core challenge in fine-grained surgical action detection, which standard object detection frameworks like Faster R-CNN [178] are not designed to handle, is the semantic ambiguity of instrument-tissue interactions. In the surgical context, the mere presence or location of an instrument is insufficient; the action is defined by the subtle interplay between the instrument (the "anchor") and its immediate spatial and temporal context. For instance, the same instrument tip (e.g., a phaco probe) can perform distinct actions like "aspirate lens nucleus" or "push lens nucleus," which are visually similar in a static frame but are defined by their motion and interaction with the surrounding tissue (the "context").

Furthermore, as noted in Chapter 1 (Section 1.2.2.1), surgical videos are fraught with visual challenges such as occlusion and motion blur, which can degrade feature quality. A robust method must therefore move beyond simple bounding box classification and instead learn a relational model. This motivates our design of the Anchor-Context Detection (ACD) module. We conceptualize the instrument as the "anchor" for the action, as it is the primary agent. The purpose of the Spatio-Temporal Action Interaction Block (STAB) is to explicitly model the relationship between this anchor and its surrounding spatio-temporal context, thereby learning a discriminative feature representation that can disambiguate fine-grained actions. The structure of STAB is shown in Figure 4.1. We introduce spatial and temporal interactions respectively in the following.

For spatial interaction: The instrument feature i_t acts as the anchor. In order to im-

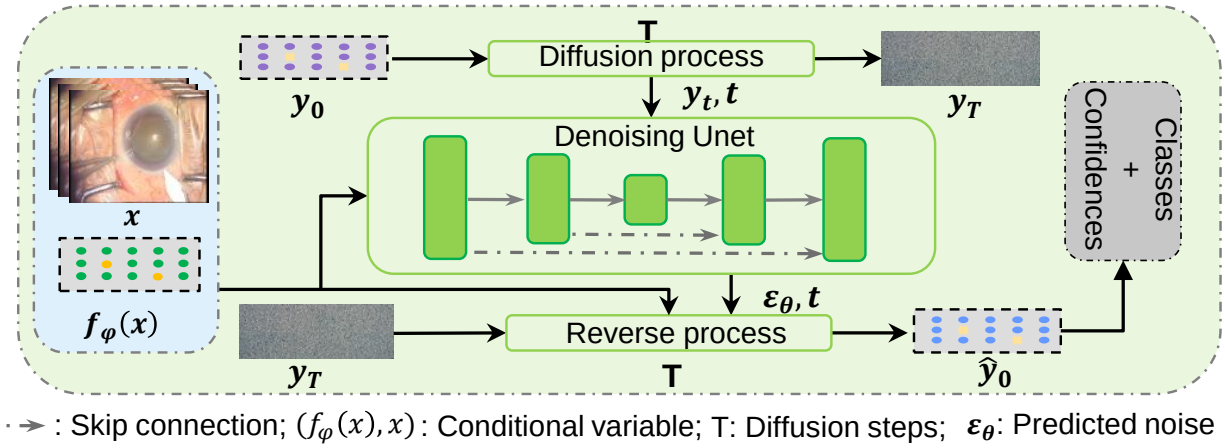


Figure 4.2: Overview of the Class Conditional Diffusion (CCD) model.

prove the anchor features, the module has the ability to select value features that are highly active with the anchor features and merge them. The formulation is defined as: $a_t = \frac{1}{C(f_t)} \sum_{j \in S_j} h(f_{tj}, i_t) g(f_{tj})$, where j is the index that enumerates all possible positions of f_t. A pairwise function $h(\cdot)$ computes the relationship such as affinity between i_t and all f_{tj}. In this work, dot-product is employed to compute the similarity. The unary function $g(f_{tj})$ computes a representation of the input signal at the position j. The response is normalized by a factor $C(f_t) = \sum_{j \in S_j} h(f_{tj}, i_t)$. S_j represents the set of all positions j. Through the formulation, the output a_t obtains more information from the positions related to the instrument and catches interactions in space for the actions.

For temporal interaction: We build memory features consisting of features in consecutive frames: $m_t = [f_{t-L}, ..., f_{t-1}]$. To effectively model temporal interactions of the anchor, the network offers a powerful tool for capturing the complex and dynamic dependencies that exist between elements in sequential data and anchors. Same with the spatial interaction, we take i_t as an anchor and calculate the interactions between the memory features and the anchor. The formulation is defined as: $b_t = \frac{1}{C(m_t)} \sum_{j \in T_j} h(m_{tj}, i_t) g(m_{tj})$,where T_j refers to the set of all possible positions along the time series in the range of L. In this way, temporal interactions with anchors are obtained. Then a global average pooling is carried out on the

spatial and temporal outputs. Action localizations and initial action class distributions are produced based on the fully-connected classifier layer.

4.2.2 Conditional Diffusion for Action Detection

While the ACD module provides a robust discriminative prediction of action location and class, it shares a limitation with most detection models: it provides a point estimate (the most likely class) without an intrinsic measure of its own certainty. As discussed in Section 4.1, this lack of reliable confidence scoring is a major barrier to clinical trust and adoption [171, 172].

To address this gap, we introduce a second-stage generative module, the Class Conditional Diffusion (CCD) model. The motivation for using a diffusion model is twofold. First, it allows us to re-frame the classification task as a generative "denoising" process. The ACD module's output can be treated as a "noisy" estimate of the true action distribution. The CCD module, trained on the full distribution of action classes, learns the complex priors of the surgical procedure (e.g., which actions are common, which are rare, which co-occur). It can then refine the initial prediction by "denoising" it towards a more plausible state within this learned distribution.

Second, and most critically, the inherent stochastic nature of the diffusion process provides a principled mechanism for uncertainty quantification. Conventional classification models typically provide a single point estimate (e.g., a softmax probability), which is often poorly calibrated and not a true measure of model confidence. In contrast, the diffusion model's generative nature allows us to sample the reverse process multiple times, generating a distribution of possible predictions. The variance of this distribution (which we term "Interval Width" or IW) serves as a direct, quantifiable measure of the model's certainty, a metric

unavailable in standard deterministic approaches. A narrow distribution (low IW) signifies high confidence, whereas a wide distribution (high IW) indicates high uncertainty, effectively flagging the prediction as unreliable.

To implement this, we adapt the continuous-space formulation of Denoising Diffusion Probabilistic Models (DDPMs) [179] to our classification task. Let $y_0 \in \mathbb{R}^n$ be a sample from our data distribution. As shown in 4.2 A diffusion model specified in continuous time is a generative model with latent $y_t = \{y_t | t \in [\lambda_{min}, \lambda_{max}]\}$ for hyperparameters $\lambda_{min} < \lambda_{max} \in \mathbb{R}$, obeying a forward process $q_t(y_t|y_{t-1})$ starting at data y_0 [179]. In this work, y_0 indicates a one-hot encoded label vector. While diffusion models typically operate on continuous data (like images), their application to classification is achieved by a conceptual shift: we do not add noise to a discrete label, but rather to a continuous vector embedding that represents that label (a "class prototype"). This construction treats the classification problem as a generative one, where the goal is to reconstruct the correct class prototype from noise. This continuous-space assumption allows us to leverage the standard Gaussian diffusion framework [180]. The primary motivation for this approach is that the stochastic, generative nature of the reverse process allows for sampling multiple predictions, which is the mechanism that enables robust confidence estimation (detailed in Section 4.3.5). The forward process is a Gaussian process that satisfies the Markovian structure:

$$q_t(y_t|y_{t-1}) = N(y_t; \sqrt{1-\beta_t}y_{t-1}, \beta_t I) \tag{4.1}$$

Furthermore, y_t can be sampled directly from y_0 using the following accumulated kernel:

$$\begin{aligned} q_t(y_t|y_0) &= N(y_t; \sqrt{\bar{\alpha}_t}y_0, (1-\bar{\alpha}_t)I) \\ y_t &= \sqrt{\bar{\alpha}_t}y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \end{aligned} \tag{4.2}$$

where $\alpha_t := 1 - \beta_t, \bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$, and $\epsilon \sim N(0,1)$. The generation of new samples can be achieved through the reversal of the forward diffusion process, followed by solving the reverse diffusion process with the Gaussian noise y_t as the initial input. The reverse diffusion

process can be computed using the following transition kernel [180]:

$$\begin{aligned} p_t(y_{t-1}|y_t, y_0) &= N\left(y_{t-1}; \tilde{\mu}_t(y_t, y_0), \tilde{\beta}_t I\right) \\ \tilde{\mu}_t(y_t, y_0) &:= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} y_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} y_t \\ \tilde{\beta}_t &:= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \end{aligned} \tag{4.3}$$

Since y_0 given y_t is unknown, it can be estimated using equation (4.2). This allows us to reverse the process from noise to data. The loss function of the neural network is:

$$L(\theta) = E_{t,y_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right] \tag{4.4}$$

where $\epsilon_\theta(y_t|t)$ estimates ϵ using a time-conditional neural network parameterized by θ.

Here, for the diffusion model optimization to be better guided by meaningful information, we integrate the action detection model as a prior or constraint in the diffusion training process. We design a conditional diffusion model $\hat{p}_\theta(y_{t-1}|y_t, x)$ that is conditioned on an additional latent variable x. In our work, the x indicates the input surgical video frames. Specifically, the model $\hat{p}_\theta(y_{t-1}|y_t, x)$ is built to approximate the corresponding tractable ground-truth denoising transition step $\hat{p}_t(y_{t-1}|y_t, y_0, x)$. we specify the forward process conditional distributions in a similar fashion as [181]:

$$\begin{aligned} &\hat{p}_t\left(y_{t-1}|y_t, y_0, x\right) = \hat{p}_t\left(y_{t-1}|y_t, y_0, f_\varphi(x)\right) = N\left(y_{t-1}; \hat{\mu}\left(y_t, y_0, f_\varphi(x)\right), \hat{\beta}_t I\right) \\ &\hat{\mu} := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} y_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} y_t + \left(1 + \frac{(\sqrt{\bar{\alpha}_t}-1)(\sqrt{\alpha_t}+\sqrt{\bar{\alpha}_{t-1}})}{1-\bar{\alpha}_t}\right) f_\varphi(x) \end{aligned} \tag{4.5}$$

where $\hat{\beta}_t = \tilde{\beta}_t$, and $f_\varphi(x)$ is the prior knowledge of the relation between x and y_0, e.g., an action detection network pre-trained with our surgical video dataset, or 0 if the relation is unknown. Consequently, the corresponding loss function of the neural network for estimating ϵ_θ is:

$$\hat{L}(\theta) = E_{t,y_0,\epsilon,x}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon + (1-\sqrt{\bar{\alpha}_t}) f_\varphi(x), x, f_\varphi(x), t)\|^2\right] \tag{4.6}$$

To access model confidence for each test instance, we mainly calculate the prediction IW. Specifically, we first sample N class prototype reconstruction with the trained diffusion model. Then calculate the IW between the 2.5^{th} and 97.5^{th} percentiles of the N reconstructed values for all test classes. Compared with traditional classifiers to precisely reconstruct the original prototype vector without much uncertainty, the denoising diffusion model is a preferable modelling choice due to its ability to produce stochastic outputs, which enable confidence generation.

4.3 Experimental Results

4.3.1 Cataract Surgical Video Dataset

The experiments for fine-grained action detection in this chapter utilize the action-annotated subset of the multi-level cataract dataset introduced in Chapter 3 (see Section 3.1.4 for details). This subset consists of 20 videos with a frame rate of 1 fps (a total of 17511 frames and 28426 action instances). Under the direction of ophthalmologists, each video is labelled frame by frame with the categories and locations of the actions. 49 types of action bounding boxes as well as class labels are included in our dataset. The surgical video dataset is randomly split into a training set with 15 videos (13583 frames) and a testing set with 5 videos (3928 frames).

4.3.2 Implementation Details

The proposed architecture is implemented using the publicly available Pytorch Library [182]. For our action detection model, in the training phase, we utilize stochastic gradient descent to optimize the model (100 epochs), with a gradually decreasing learning rate starting from

Table 4.1: Methods comparison and ablation study on cataract video dataset.

Methods	mAP_{10}	mAP_{30}	mAP_{50}	mAP_{mean}
Faster R-CNN [178]	0.388	0.384	0.371	0.381
SSD [183]	0.360	0.358	0.350	0.356
RetinaNet [184]	0.358	0.356	0.347	0.354
Mask R-CNN [185]	0.375	0.373	0.363	0.370
SSDlite [186]	0.305	0.304	0.298	0.302
Dynamic R-CNN [187]	0.315	0.310	0.296	0.307
OA-MIL [188]	0.395	0.394	0.378	0.389
backbone	0.373	0.365	0.360	0.366
+temporal	0.385	0.378	0.372	0.378
+spatial	0.394	0.385	0.377	0.385
+STAB	0.400	0.393	0.385	0.393
+CCD (ACTNet)	**0.415**	**0.406**	**0.397**	**0.406**

0.01, a momentum of 0.9 and a batch size of 16. For the diffusion model, we employ an Adam optimizer to optimize the model (300 epochs), with a learning rate of 0.001 and a batch size of 16. For all the processes of training and testing, we conduct on one NVIDIA RTX A6000. A model with ResNet50 backbone from Faster R-CNN-benchmark [178] is adopted for our instrument anchor detection. During inference, detected anchor boxes with a confidence score larger than 0.8 are used. The performances are evaluated with official metric frame level mean Average Precision (mAP) at IoU = 0.1, 0.3 and 0.5, respectively, obtaining figures we call in the following mAP_{10}, mAP_{30} and mAP_{50}. Their mean mAP_{mean} is also computed to get a final evaluation score.

4.3.3 Method Comparison

In order to demonstrate the superiority of the proposed method for surgical action detection, we carry out a comprehensive comparison between the proposed method and the following representative methods: 1) single-stage algorithms, including the Single Shot Detector (SSD) [183], SSDLite [186] and RetinaNet [184]. 2) two-stage algorithms, including Faster R-CNN [178], Mask R-CNN [185], Dynamic R-CNN [187] and OA-MIL [188]. The data presented in Table 4.1 clearly demonstrate that our method outperforms other approaches, irrespective of the IoU threshold being set to 0.1, 0.3, 0.5, or the average values. Notably, the results obtained after incorporating diffusion even surpass Faster R-CNN by 2.5% and baseline by 4.0% in terms of average mAP. This finding provides compelling evidence for the efficacy of our method. The 4.0% absolute improvement in $\boldsymbol{mAP}_{mean}$ over the baseline (backbone) and 2.5% over the strong Faster R-CNN baseline are statistically significant margins, underscoring the value of integrating spatio-temporal interactive information (via STAB) and leveraging a generative diffusion model (via CCD) to optimize the category distribution.

The quantitative results further corroborate the effectiveness of our approach in Figure 4.3, which shows that our model does not only improve the performance of the baseline models but also localizes accurately the regions of interest of the actions. Figure 4.4 lists more quantitative results on our cataract dataset. We add the comparison of our ACD method. It can be seen that our method can accurately identify action categories with precise localization, and give prediction confidence.

4.3.4 Ablation Study

To validate the effectiveness of our ACTNet, we performed a detailed ablation study, with results presented in the lower half of Table 4.1.

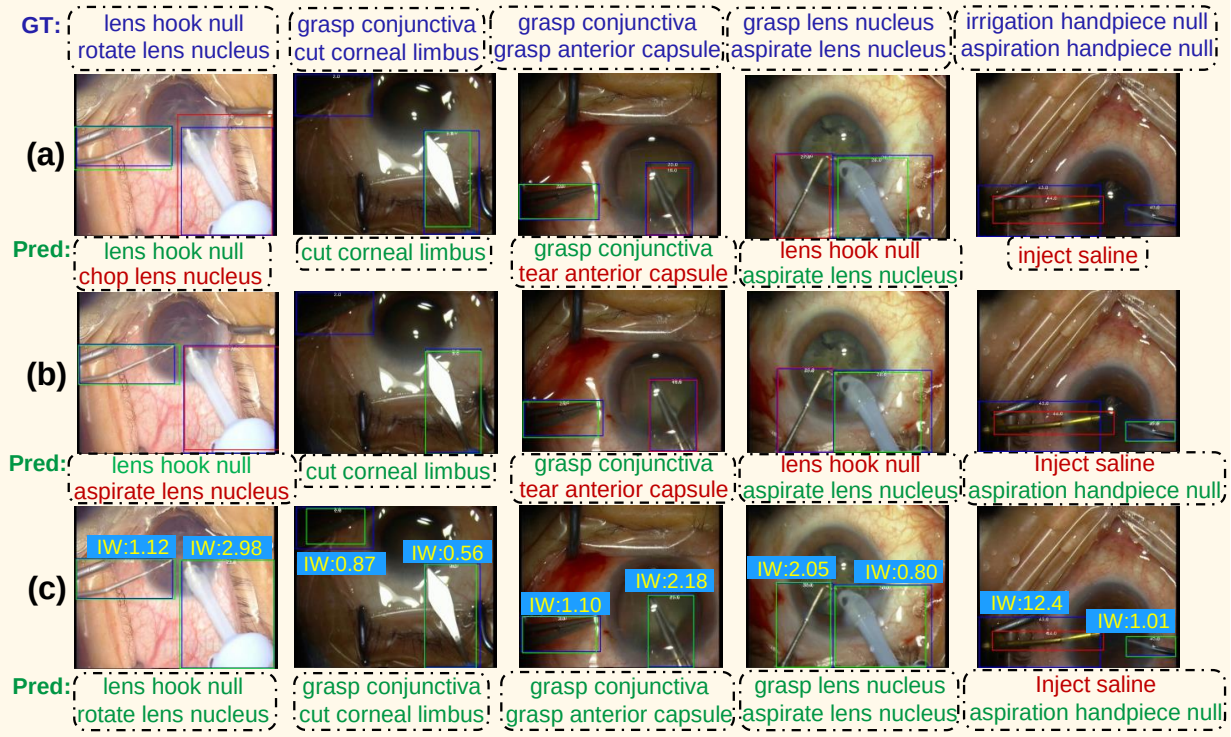


Figure 4.3: Visualization on cataract dataset. We choose different actions to show the results of (a) Faster R-CNN, (b) OA_MIL, and (c) our ACTNet. For each example, we show the ground-truth (Blue), right predictions (Green) and wrong predictions (Red). The actions are labelled from left to right. IW values (multiplied by 100) mean prediction interval width to show the level of confidence.

The "backbone" model, which serves as our baseline, yields a mAP_{mean} of 0.366. This model concatenates anchor features (from RoIAlign) with the corresponding general frame features. Its modest performance highlights the difficulty of the task when relational information is not explicitly modeled.

We then isolated the contributions of spatial and temporal context. Adding only temporal context ("+temporal") or only spatial context ("+spatial") alone provides a clear performance boost, raising the mAP_{mean} to 0.378 and 0.385, respectively. This confirms that both temporal motion and spatial instrument-tissue interaction are key signals for this task. Notably, spatial interaction provides a slightly larger benefit, suggesting that the precise

instrument-tissue relationship is a highly discriminative feature.

Crucially, the full STAB module ("+STAB"), which integrates both spatio-temporal interactions as described in Section 4.2.1.2, achieves a mAP_{mean} of 0.393. This synergistic effect, yielding a 2.7% gain over the backbone (and outperforming the simple sum of the individual temporal and spatial gains), confirms our core hypothesis: surgical action is not just a spatial or a temporal phenomenon, but a truly spatio-temporal one that benefits from joint modeling.

Finally, the addition of the CCD module ("+CCD (ACTNet)") provides a further 1.3% gain, pushing the final mAP_{mean} to 0.406. It is important to note that the mAP metric incorporates both localization (IoU) and classification accuracy. The CCD module operates purely on the classification component; it does not alter the bounding box locations proposed by the ACD module. This 1.3% gain is therefore attributed entirely to improved classification accuracy. The CCD module corrects classification errors made by the ACD module (e.g., turning false positives with correct location but wrong class into true positives). This demonstrates the value of the generative refinement stage, which leverages the learned priors of the surgical workflow to clean up classification ambiguities in the ACD module's predictions.

4.3.5 Confidence Analysis

To analyze the model confidence generated by our CCD module, we take the best prediction for each instance to calculate the instance accuracy. As detailed in Section 4.2.2, we quantify this confidence by sampling the reverse diffusion process multiple times ($N = 100$) and calculating the IW of the resulting prediction distribution. We can observe from Table 4.2 across the test set, the mean IW of the class label among correctly classified instances by

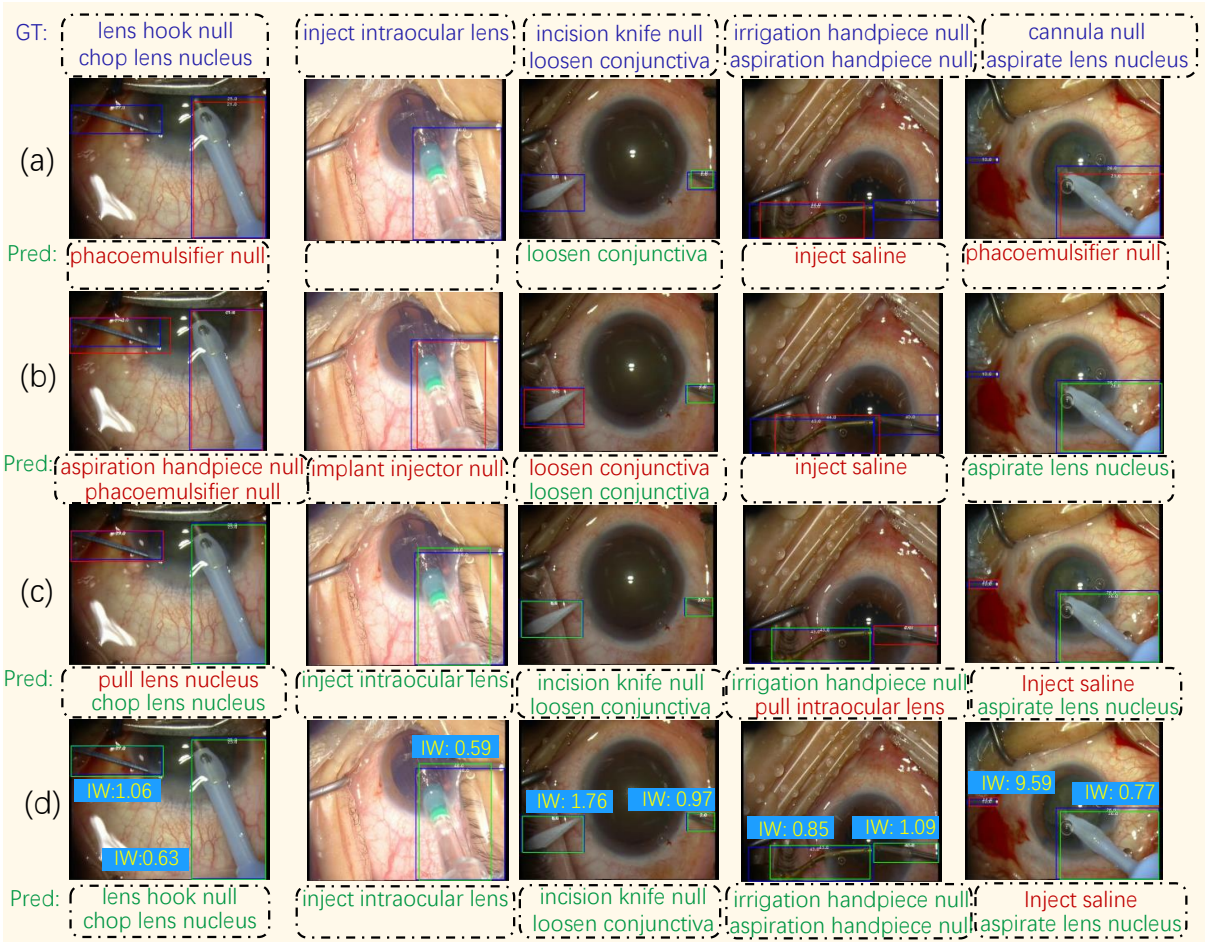


Figure 4.4: Visualization on cataract dataset. We choose different actions to show the results of (a) Faster R-CNN, (b) OA_MIL, (c) our ACD and (d) our ACTNet. For each example, we show the ground-truth (Blue), right predictions (Green) and wrong predictions (Red). The actions are labelled from left to right. IW values indicate prediction Interval Width to show the level of confidence.

ACTNet is significantly narrower compared to that of incorrectly classified instances. This observation indicates that the model is more confident in its accurate predictions and is more likely to make errors when its predictions are vague.

The "significance" of this result, as seen in Table 4.2, is the large, consistent gap between the Mean_IW for correct (e.g., 0.91, 0.54) and incorrect (e.g., 9.78, 9.73) predictions. This wide separation demonstrates that the IW is a reliable proxy for model uncertainty: when

the model is correct, its generative process repeatedly converges to the same stable prototype (narrow IW); when it is incorrect, the model is uncertain, and its stochastic sampling results in a wider distribution of possible predictions (wide IW), effectively signaling its own lack of confidence.

The practical significance of this finding, as detailed in Table 4.2, is substantial for clinical translation. This analysis moves beyond simple accuracy metrics to evaluate the model's self-awareness, a key requirement for safety-critical systems [171, 172]. For instance, in the "grasp conjunctiva" class, the mean IW for correctly classified instances is 0.91. In contrast, for incorrectly classified instances of the same class, the mean IW is 9.78—an order of magnitude larger. This wide gap demonstrates the model's ability to "know what it doesn't know". A high IW value, such as 9.78, can serve as a direct, actionable flag to a Context-Aware System (CAS). This flag indicates that the model's prediction is highly unreliable and should either be suppressed from the surgeon's view or presented with a low-confidence warning.

This capability is observed across all classes. For "aspirate lens nucleus," the IW for correct predictions is 0.76, versus 10.30 for incorrect ones. This robust correlation between low confidence (high IW) and incorrect predictions is precisely the feature required for building trust. It allows the system to fail safely and predictably, rather than failing silently. This mechanism is a foundational step toward integrating deep learning-based recognition into real-time clinical decision support, as it provides the reliability metric that has been largely missing from previous surgical action detection models. Furthermore, upon comparing the mean_IW at the true class level, we find that a more precise class tends to exhibit a larger disparity between the correct and incorrect predictions. Figure 4.3 also shows the confidence estimations for some samples. We can see the correct prediction gets smaller IW values compared with the incorrect one (The rightmost figure in column (c)), which means it has more uncertainty for the incorrect prediction.

Table 4.2: The mean_IW (multiplied by 100) results from our CCD module.

Class	Instance	Accuracy	Mean_IW (Correct)	Mean_IW (Incorrect)
grasp conjunctiva	487	0.702	0.91 (342)	9.78 (145)
aspirate lens cortex	168	0.613	1.37 (103)	15.82 (65)
chop lens nucleus	652	0.607	0.54 (396)	9.73 (256)
polish intraocular lens	222	0.572	0.90 (127)	8.48 (95)
aspirate lens nucleus	621	0.554	0.76 (344)	10.30 (277)
inject viscoelastic	112	0.536	2.17 (60)	9.14 (52)
Remove lens cortex	174	0.471	0.42 (82)	5.84 (92)
forceps null	280	0.464	2.67 (130)	8.38 (150)

4.3.6 Qualitative Analysis and Limitations

While the quantitative results in Table 4.1 demonstrate the effectiveness of ACTNet, a qualitative analysis of its failure modes reveals key limitations that motivate the subsequent chapters of this chapter. The primary sources of error were identified as:

• **Visual Ambiguity:** The model still struggles with actions that are visually similar and differ only by subtle context or lack of motion. For example, as seen in the rightmost column of Figure 4.3 , the ground truth is irrigation handpiece null (an idle state). However, ACTNet, along with other baseline models, misclassifies this as Inject saline. Although our CCD module correctly assigned a very high Interval Width (IW: 12.4) to this prediction, indicating high uncertainty, the underlying classification is still incorrect. This failure highlights the difficulty of distinguishing an idle instrument from an active one (like injecting clear fluid) using local spatio-temporal features alone.

• **Anchor Occlusion or Blur:** The ACTNet framework relies heavily on the "anchor"

(the instrument feature). In frames where the instrument tip is heavily occluded by tissue or suffers from significant motion blur (as discussed in Section 1.2.2.1), the ACD module fails to extract a discriminative anchor feature. This leads to missed detections or classifications.

These limitations strongly motivate the hierarchical framework in Chapter 5, which uses high-level phase and step context to constrain the predictions of low-level actions (e.g., knowing the phase is not "Incision" might reduce the likelihood of an "Inject saline" prediction). They also motivate the causality-inspired framework in Chapter 6, which explicitly treats visual ambiguities as confounders to be mitigated.

4.4 Summary

In this chapter, we propose a conditional diffusion-based Anchor-context spatio-temporal Action Detection Network (ACTNet) to achieve recognition and localization of every occurring action in the surgical scenes. ACTNet improves the accuracy of the predicted action class from two considerations, including spatio-temporal interactions with anchors by the proposed Spatio-Temporal Anchor interaction Block (STAB) and full distribution of action classes by Class Conditional Diffusion (CCD) module, which also provides uncertainty in surgical scenes. Experiments based on cataract surgery demonstrate the effectiveness of our method. Overall, the proposed ACTNet presents a promising avenue for improving the accuracy and reliability of action detection in surgical scenes.

Chapter Five

Multi-level Surgical Video Scene Understanding

This chapter is based on the following publications:

Hierarchical Context Transformer for Multi-level Semantic Scene Understanding [189], which is published in IEEE Transactions on Circuits and Systems for Video Technology (2025).

Chapter 4 introduced ACTNet, a framework to address the challenge of fine-grained action detection, successfully improving localization and classification accuracy. However, that work focused on only one level of the surgical hierarchy. It did not explicitly model the rich, contextual relationships between fine-grained actions and the broader surgical phases and steps. As outlined in Chapter 1, a truly holistic understanding requires a framework that can reason across all levels simultaneously. This chapter directly addresses that limitation by proposing a Hierarchical Context Transformer designed for multi-level semantic scene understanding.

5.1 Introduction

Computer Assisted Surgical (CAS) systems now have a vital role in enhancing the quality of interventional healthcare [12], surgeon training as well as assist procedure planning and retrospective analysis [163, 166]. To achieve it, CAS systems need to make use of a complete and explicit understanding of surgical scenes [45, 166, 190, 191]. Specifically, surgical phases and steps can monitor surgical processes and provide early alerts of potential forthcoming anomalies [192, 193]. Detailed instruments and actions appearing in the surgical scenes are beneficial to reducing preventable operation errors [194, 195], clinical decision support [167] and skill assessment [196–199].

A surgical procedure can be decomposed at different levels of granularity, such as the whole procedure, phases, steps, and actions [166].

Most of the currently available methods for surgical understanding are only focused on one specific level of granularity, like phase or step recognition at coarse granularity [40, 64, 92, 95, 200–203] or action recognition at fine granularity [44, 102, 204–206], as shown in Figure 5.1 (A), (B), (C) respectively. Nevertheless, a comprehensive surgical scene understanding plays

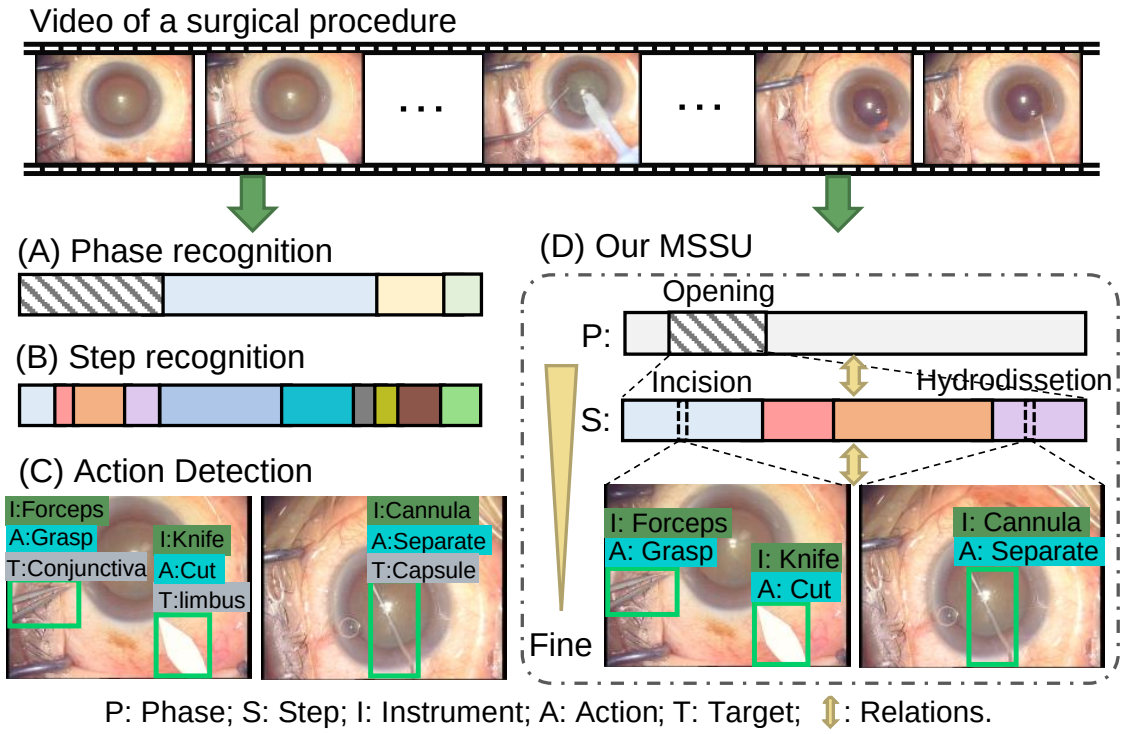


Figure 5.1: Comparison of our proposed method with existing methods. MSSU: Multi-level Semantic Scene Understanding in surgery.

a key role in developing CAS systems, which needs to be started from a global overview of a description of high-level tasks to fine details [166]. Specifically, phases and steps as coarse parts provide general information about the ongoing surgical procedure, while fine-grained action and instrument detection provide a detailed perception and understanding of surgical scenes, as shown in Figure 5.1 (D). Semantic information from different levels can complement each other, forming a complete surgical portrait from various views. However, few works provide systematical analysis across all the levels of surgical interpretations, and those that do often fail to explore the relationships between different levels [21]. More critically, different levels of representation facilitate each other by utilizing the containment information from these hierarchies. For example, in cataract surgery, as shown in Figure 5.1 (D), the instrument "incision knife" is often used to act "cut corneal-limbus", and it generally appears in phase "opening" and step "incision". Similarly, step "hydrodissetion"

often appears during phase "opening" and also indicates the occurrence of both instrument "cannula" and action "separate capsule". Stressing on their intrinsic complementarity, the relationship between multi-level activities is essential for understanding the scene dynamics [21, 207]. Motivated by this, we propose a hierarchical context network for Multi-level Semantic Scene Understanding (MSSU) (phase recognition → step recognition → action and instrument detection) and explore the relationship among them.

To fully utilize the cross-task knowledge, we design an inter-task contrastive learning algorithm to further boost our model performance. Specifically, contrastive learning can automatically learn the semantic alignment relation across different tasks according to the actual data distribution [208]. In this work, we treat the inter-task representations of the same scenes as positive, while the representations of other random scenes as negative. Based on this, the contrastive learning approach is capable of maximizing the mutual information between different views of the same scene, resulting in more compact and better feature representations.

Furthermore, vision transformers have shown outstanding capabilities in visual feature representation. Nonetheless, the growing number of parameters and model size pose limitations on training from scratch and the extensive adoption of transformers, particularly when training video models in surgical scenes. Parameter-efficient transfer learning is an excellent way to overcome this drawback [209, 210]. Besides, inferring spatial and temporal structured information is a crucial aspect in the context of surgical video understanding. Therefore, with the above observation, we propose a simple Spatial-Temporal Adapter (ST-Ada) capable of learning the representations in space and time to achieve effective video understanding in surgical scenes at a small parameter cost.

In this chapter, we present a Hierarchical Context Transformer (HCT) network to achieve MSSU in surgical videos. To realise this, we first design a Hierarchical Relation Aggrega-

tion Module (HRAM) to generate hierarchical representations for different tasks and further capture the relations between them. Then, after fusing the cross-task information guided by correlations, we obtain task-specific representations. Next, to enhance the consistency of inherently existing crossing the different tasks, we propose Inter-task Contrastive Learning (ICL) to make the task-specific representation more discriminative. Considering the computation cost and model storage when transferring knowledge from a pre-trained model in transformer architectures, we present HCT+ network to add ST-Ada to the HCT to effectively learn spatial, and more importantly temporal representations with much fewer parameters for our multi-task learning at a significantly smaller computational cost.

In summary, the key contributions of this work are four-fold:

- We propose a novel Hierarchical Context Transformer (HCT) network to achieve the Multi-level Semantic Scene Understanding (MSSU) and explore inter-relations between different tasks fully. MSSU represents the recognition and detection of phase → step → action and instruments in surgery.
- We propose a Hierarchical Relation Aggregation Module (HRAM) to encode the phase-step-instrument-action hierarchical relationship. Furthermore, for any given task, HRAM possesses the capability to aggregate information contributed by the other three tasks and exploit it to enhance task-wise representation learning.
- We develop an Inter-task Contrastive Learning (ICL) to further strengthen the context learning capability of the model by the inherent consistency across the tasks. Considering the high computational costs of the transformer model, we integrate a Spatial-Temporal Adapter (ST-Ada) into our HCT denoted as HCT+, which attains strong transfer learning abilities in spatial and temporal representations at a small parameter cost.

- We conduct experiments with the proposed network on four datasets, our cataract video dataset, the public endoscopic dataset PSI-AVA and MultiBypass140, and even the natural dataset 50Salads, to demonstrate its State-Of-The-Art (SOTA) performance and the contribution of each of our proposed components.

5.2 Related work

5.2.1 Surgical Scene Understanding

Recent works in surgical scene understanding can be broadly classified into two categories. One category aims to automatically identify surgical phases or steps within untrimmed surgical videos, which has received a lot of attention and is a very active area of research. TMRNet [200] is proposed to establish a long-range memory bank to provide long-term temporal features and improve phase recognition accuracy. Yi et al. [201] proposed a non-end-to-end training strategy based on MS-TCN. After the transformer gained significant popularity for processing sequential data, Pan et al. [202] exploited the Swin transformer and LSTM networks to recognise phases in surgical videos. Czempiel et al. devised OperA [92], a transformer-based model, to extract high-quality frames with an attention regularization loss. Ding et al. [95] presented SAHC for surgical phase recognition by emphasizing learning segment-level semantics to address erroneous predictions caused by ambiguous frames. Yue et al. [203] fused the frame-level and phase-level temporal context for surgical workflow analysis.

Another category focuses on finer-grained action recognition. Bawa et al. [97] presented an endoscopic minimally invasive surgery dataset designed to tackle the problem of surgeon action detection, and provided baseline results with open-source methods like [98]. Hao et al.

[163] presented ACTNet to get action locations, classes and confidence guided by anchors. Unlike existing methods for a single task, we design a new transformer module to learn the relations from multi-level tasks. Besides, we also model the constraints in the hierarchical information to regularize the framework with the ICL comprehensively.

5.2.2 Multi-task Learning in Surgical Scene

For improving performance, Ramesh et al. [25] proposed a multi-task MS-TCN to jointly predict the phases and steps, failing to get a complete surgical scene understanding. Nwoye et al. proposed the action triplets (instrument, verb, target, as shown in Figure 5.1 (C)) to represent the actions in a laparoscopic dataset [100], and a trainable 3D interaction space to capture the associations between the triplet components. They further developed a new model, the Rendezvous [44], to detect triplet components with class activation-guided and mixed attentions. After adding the location information of the instrument and target, Lin et al. [102, 103] presented a novel QDNet for the instrument-tissue interaction quintuple detection task in cataract videos. Valderrama et al. [21] introduced the PSI-AVA dataset and proposed a baseline framework for different level task achievement. However, they failed to harness and explore the interdependencies and correlations among these tasks. Although most of these works employ multi-task learning for action recognition, they are confined to the singular horizontal dimension of actions for more detailed analyses, while neglecting the complementarity of vertical information such as phases and steps. Our work, in contrast, encompasses an integrated analysis across both horizontal and vertical dimensions in surgical scenarios.

5.3 Methodology

5.3.1 Overview

The overall framework of HCT illustrated in Figure 5.2, for achieving multi-level semantic activity recognition in surgical scenes, from phase and step recognition to finer-grained action and instrument/tool detection denoted by p, s, a and t respectively, consists of only transformer blocks. Specifically, to generate the feature of frame x_t, we extract a video clip as the model input, which contains the current frame x_t and a set of its previous frames, as $x = x_{t-l}, ..., x_{t-1}, x_t$, with $l+1$ frames. The clip is first passed forward a transformer backbone to generate shared feature maps f_p, f_s and f_a. f_t is instrument-specific features. For obtaining video features effectively to model the complex temporal cues of the surgical scenes, we utilize the MViTv2 [211] model as the backbone. It allows capturing simple low-level visual information at early layers and spatially coarse, but complex and high-dimensional features at deeper layers, benefiting from its multi-scale feature hierarchies.

After that, an HRAM is proposed to aggregate different tasks to dig out the dependencies among them and further get the task-specific features. Specifically, we utilize cross-attention in transformer blocks to get the relations $T_{i \leftrightarrow j}$ ($i, j \in [p, s, a, t]$ and $i \neq j$) cross the different task i and j. Then the relations are fused with corresponding self-attention task-wise features to get $\dot{f}_p$, $\dot{f}_s$ and $\dot{f}_a$. To better preserve the information of each task, we add skip connections as illustrated in Figure 5.2 to get $\ddot{f}_p$, $\ddot{f}_s$ and $\ddot{f}_a$. Then, to acquire a more compact and better feature representation for different tasks, we develop the ICL approach to maximize the mutual information between different views of the same scenes. In this work, $p \leftrightarrow s$ and $a \leftrightarrow t$ are selected as two contrastive pairs because phase and step have strong consistencies for the same positive sample, as do action and instrument, which is also validated by our experiments. The HCT is also optimised by the supervision losses for each task $\mathcal{L}_p$, $\mathcal{L}_s$, $\mathcal{L}_a$

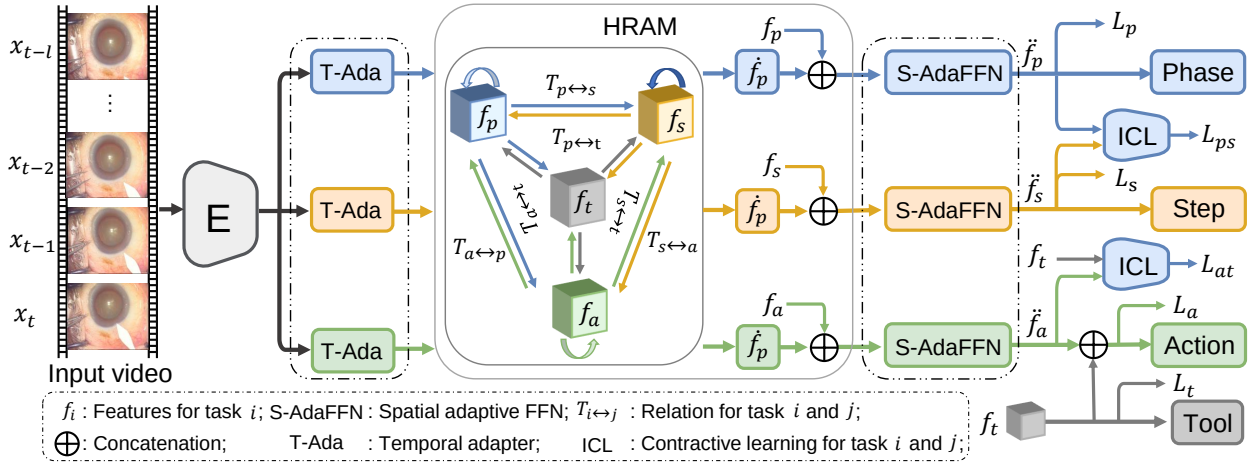


Figure 5.2: The pipeline of our Hierarchical Context Transformer (HCT) framework. Given an input video with $l+1$ frames, HCT uses a transformer model to extract shared features, which are then fed into the Hierarchical Relation Aggregation Module (HRAM) to capture the relations between the four task-wise features. After that, Inter-task Contractive Learning (ICL) is utilized to further optimize the HCT. For the plus version of the model HCT+, we add a temporal adapter before HRAM and put the spatial adapter in the feed-forward module.

and $\mathcal{L}_t$.

Furthermore, the extensive parameter size of the transformer model imposes significant limitations on its utility, particularly in the medical scenes. With this in mind, this chapter introduces the temporal adapter (T-Ada in Figure 5.2) placed before the HRAM and Spatial Adapter placed in the Feed-Forward Network (S-AdaFFN in Figure 5.2) to construct HCT+. It attains strong transfer learning abilities by only fine-tuning a small number of extra parameters while getting competitive performance.

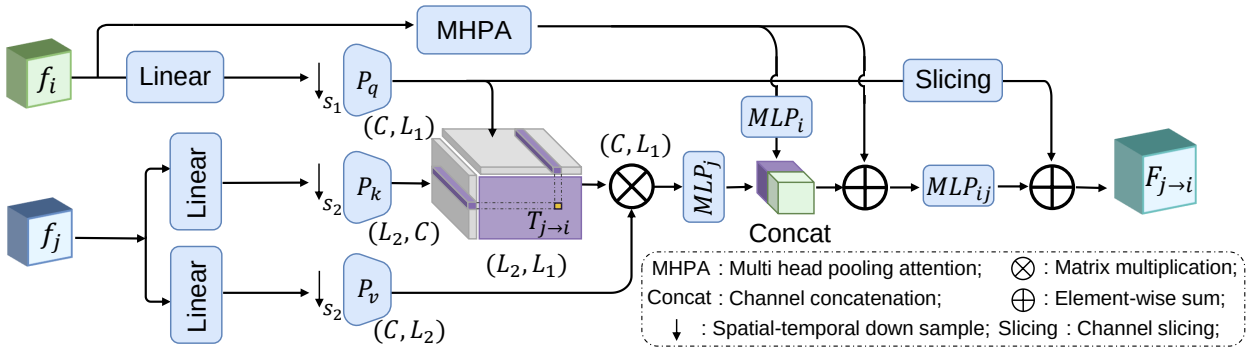


Figure 5.3: Detailed structure of the Hierarchical Relation Aggregation Module (HRAM).

5.3.2 Hierarchical Relation Aggregation Module (HRAM)

Transformer has strength in the computation of long-range dependencies and the general representation ability for various tasks [95, 202]. This advantage is essential since the single branches start from the shared features. In this work, we achieve the shared encoder built based on a stack of transformer blocks from MViTv2. After getting shared features, in order to exploit the hierarchical relationship in the surgery, HRAM is developed to generate task-specific features with complementary information inside.

As illustrated in Figure 5.3, suppose there are n tasks. The input features of the HRAM are denoted as f_i and f_j, where $i, j \in \{1, ..., n\}$ and $i \neq j$. Without loss of generality, considering here task i as our primary task, and task j as the secondary, we seek to estimate the features $f_{j\to i}$ from task j that can contribute to task i. We employ two parts of attention: (i) correlation attention, which aims to exploit cross-task spatial-temporal correlations to guide the extraction of contributing features of the secondary task to the primary task; (ii) a self-attention on the primary task, to retain and further enhance information from the primary task.

Concretely, in correlation attention part, for $f_i, f_j \in \mathbb{R}^{L\times C}$ ($L = l \times h \times w$, l denotes length of input video clip, h and w are the height and width of feature map respectively), it applies

pooling operations to perform spatial-temporal downsampling with s_1 and s_2 scale factor respectively, to get query f_q, key f_k and value f_v tensors [211].

$$f_q = P_q(f_i W_q; s_1) \quad f_k = P_k(f_j W_k; s_2) \quad f_v = P_v(f_j W_v; s_2), \tag{5.1}$$

where $f_q \in \mathbb{R}^{L_1 \times C}$ and $f_k, f_v \in \mathbb{R}^{L_2 \times C}$. W_q, W_k, W_v represent the projection matrices. Spatial-temporal correlation attention matrix $C_{j \to i} \in \mathbb{R}^{L_1 \times C}$ is then obtained :

$$C_{j \to i} = softmax(\frac{f_q f_k^T}{\sqrt{d}}) f_v, \tag{5.2}$$

where $\sqrt{d}$ is used to normalize the inner product matrix row-wise. Intuitively, $C_{j \to i}$ has high values where features from task i and j are highly correlated and low values otherwise. For the self-attention part on primary task i, Multi-Head Pooling Attention (MHPA) [212] is employed to enable progressive change at the spatial-temporal resolution in the transformer backbone.

Subsequently, n mapping functions are utilized to adjust the output dimension and fuse different parts of attention, the integrated feature $T_{j \to i}$ for task i, enhanced by the correlation between task i and j, is defined as:

$$T_{j \to i} = concat\left(MLP_j(C_{j \to i}), MLP_i(\mathrm{MHPA}(f_i))\right), \tag{5.3}$$

where MLP_j is designed to transfer the size of $C_{j \to i}$ to $L_1 \times \frac{C}{n-1}$, while MLP_i is designed to maintain the dimension of feature f_i. $concat$ indicates the channel-wise concatenation.

After that, the concatenated feature map is then processed with the element-wise addition with the self-attention information of f_i, followed by another mapping function MLP_{ij}, leading to the final refined features. Consequently, for aggregating all correlated secondary task information contributing towards primary task i, we can get final refined task-specific features $F_{all \to i}$:

$$\begin{aligned} T_{all \to i} &= MLP_{ij}\left(concat_{j=1}^{n}\left(T_{j \to i}\right) + \mathrm{MHPA}(f_i)\right), \\ F_{all \to i} &= T_{all \to i} + slicing(f_q), \end{aligned} \tag{5.4}$$

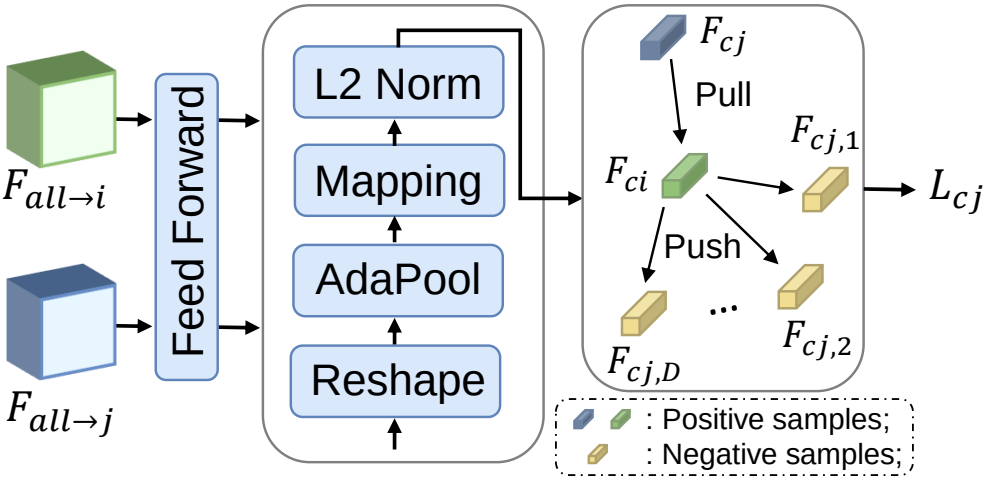


Figure 5.4: Detailed structure of the Inter-task Contrastive Learning (ICL).

where $slicing(\cdot)$ indicates channel slicing for achieving skip connection. The slicing operation shown in Figure 5.3 is optional, and is particularly useful in instrument-specific feature computation. It ensures dimensional consistency and filters out irrelevant information, which is essential for accurate processing in the task. MLP_{ij} is employed to keep the input f_i and output dimensions consistent. By design, the refined primary task features have the same dimensions as the input features.

In this chapter, we intend to achieve four tasks to attain a hierarchical understanding of surgical scenes. For the tasks of phase, step recognition and action detection, we fuse relevant information from the other three tasks and get final refined task-specific features $F_{all\to i}$. To account for instrument detection, we employ the DINO [213] to obtain bounding box estimations and box-specific features, then feed them into the HRAM module. Therefore, when we use instrument-specific features as auxiliary roles, the architecture consisting of two linear layers and an activation layer in the middle and padding operation is proposed to keep the same dimension as the primary task features to operate following attention calculation. What is more, as shown in Figure 5.2, to achieve the action detection task, we integrate action-specific features and instrument-specific features to obtain final results.

5.3.3 Inter-task Contrastive Learning (ICL)

As the different tasks are learned from the same input video clip data, consistency inherently exists crossing the different tasks in a multi-task prediction. The model performance of one task can thus benefit from utilizing the consistent information from other tasks to boost task-specific representation learning. To better illustrate such consistency, we propose ICL bring their representations closer together, which brings task pairs (positive examples) near together while pushing irrelevant ones (negative examples) far apart. The detailed structure of this ICL module is illustrated in Figure 5.4.

Specifically, with input video clips, we treat the representations from two different tasks in the same scene (same video clips) as positive pairs, while the representations in different scenes (different video clips) as negative samples. For n tasks in our paper, we take two of them as examples denoted by i and j as task pairs for contrastive learning. We set the feature obtained from the HRAM as the input of ICL. As illustrated in Figure 5.4, for the features $F_{all\to i}$ and $F_{all\to j}$, after the feed-forward process in the transformer block, we first average them in terms of the space-time dimension by adaptive average pooling, followed by a set of mapping function and L_2 normalization. Then, we get the reshaped features F_{ci} and F_{cj}. Given a positive example of task pair, we randomly pick a set of D $\{F_{cj,d}\}_{i=1}^{D}$ for F_{ci} from the same batch as negative examples and vice versa. Therefore, the losses for tasks i and j are as follows:

$$\begin{aligned}
\mathcal{L}_{ci} &= -\sum_{F_{ci},F_{cj}} log\frac{exp(sim(F_{ci},F_{cj})/\tau}{\sum_{d\in\Lambda} exp(sim(F_{ci},F_{cj,d})/\tau)},\\
\mathcal{L}_{cj} &= -\sum_{F_{ci},F_{cj}} log\frac{exp(sim(F_{ci},F_{cj})/\tau}{\sum_{d\in\Lambda} exp(sim(F_{ci,d},F_{cj})/\tau)},\\
\mathcal{L}_{cij} &= \mathcal{L}_{ci} + \mathcal{L}_{cj},
\end{aligned} \tag{5.5}$$

where Λ represents a batch of video clips, and τ is the temperature parameter that controls the strength to contrast. sim is defined as $sim(x,y) = \frac{x^T y}{|x||y|}$. In the training process, negative

examples are from the same training batch of data. $\mathcal{L}_{cij}$ denotes ICL loss between task i and task j. In experiments, phases and steps are selected to form pairs and fed into ICL due to their strong inherent consistency in the same positive samples. This holds true for pairs composed of actions and instruments as well, which is validated by subsequent ablation studies.

5.3.4 HCT+ Network

Transfer learning is an efficient way to transfer knowledge from one domain to another. However, limited by the growing number of parameters and model size, fully fine-tuning the whole transformer model for every single downstream task would become prohibitively expensive and infeasible in training cost and model storage, especially for surgical video datasets [209]. In this work, we integrate a spatial-temporal adapter into our improved transformer block, which attains strong transfer learning abilities by only fine-tuning a small number of extra parameters while achieving comparable performance with the full fine-tuned model.

Video understanding requires the model to learn both good appearance representations in each frame (spatial modeling) and also infer the temporal structured information across frames (temporal modeling). Hence, temporal and spatial adapters are employed in different places of the transformer block to model disparate information. Inspired by the Adapter [214] designed for parameter-efficient transfer learning in NLP, we adopt spatial and temporal adapters [209, 210].

As shown in Figure 5.5, before the transformer block, the input video clip features are firstly fed into the temporal adapter denoted by T-Ada, which is a bottleneck architecture that consists of two Fully Connected (FC) layers and a Depth-Wise 3D-Convolution (DW-

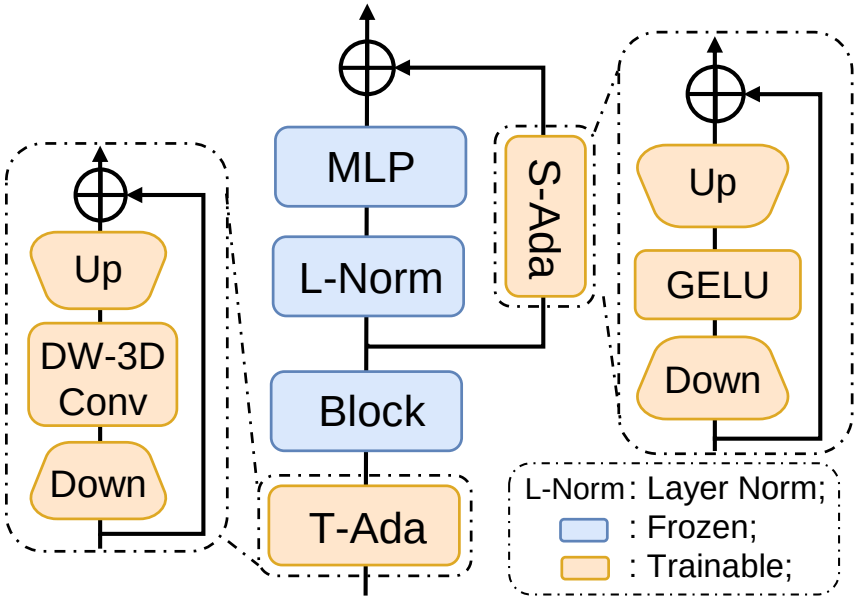


Figure 5.5: Detailed structure of the spatial-temporal adapter in our proposed transformer block.

Conv3D) in the middle [209]. The first FC layer projects the input to a lower dimension and the second FC layer projects it back to the original dimension. Formally, the temporal adapter T_Ada can be expressed as:

$$T_Ada(f_i) = f_i + \mathbf{W_{up}}\left(DW_Conv3D(\mathbf{W_{down}} f_i)\right), \tag{5.6}$$

where the $\mathbf{W_{up}} \in \mathbb{R}^{L \times \hat{L}}$ and $\mathbf{W_{down}} \in \mathbb{R}^{\hat{L} \times L}$ denote the parameters of the up-projection layer and the down-projection layer, where $\hat{L}$ is the bottleneck middle dimension and satisfies $\hat{L} << L$.

In contrast to the temporal adapter, the spatial adapter employs an intermediate activation layer GELU instead of DW-Conv3D. With the output features $F_{all \to i}$ of transformer block as input, the spatial adapter operator S_Ada is formally written as:

$$S_Ada(F_{all \to i}) = F_{all \to i} + \mathbf{W_{up}}\left(GELU(\mathbf{W_{down}} F_{all \to i})\right). \tag{5.7}$$

For proper adaptation, we add the spatial adapter to the feed-forward module to transfer

spatial information, denoted by S-Ada in Figure 5.5. During training, all the other layers of the transformer model are frozen while only the spatial and temporal adapters are updated [214].

5.3.5 Overall Training Objective

In this work, we employ a weighted binary cross-entropy loss for action detection while a weighted cross-entropy loss is utilized for the phase, step and instrument recognition. Furthermore, the ICL losses for task pairs are added to enhance the task-specific feature extraction. The final training objectives can be written as:

$$\mathcal{L}_f = \sum_{i=1}^{n} \lambda_i \mathcal{L}_i + \sum_{i,j=1,i\neq j}^{n} \mathcal{L}_{cij} \tag{5.8}$$

where λ_i, $\mathcal{L}_i$ denote the weights and loss functions for different tasks, and n is the number of tasks.

5.4 Experiments

5.4.1 Datasets

The effectiveness of our HCT is evaluated on four datasets: our cataract dataset, the public PSI-AVA [21], MultiBypass140 [47] and 50Salads [215] dataset, which are derived from distinct surgical scenes, even from natural scenes. Our private cataract dataset comprises 20 videos (a total of 17511 frames) with a frame rate of 1 fps and resolution of 720 × 576. The dataset is provided with manual annotations done by surgeons indicating the surgical phase and step each video frame belongs to, and the actions and tools appearing in the scene. Specifically, the cataract videos consist of 4 phases, 10 steps, 49 actions and 13 instruments.

For each video frame, the phase and step are provided with class labels, while the actions and instruments are labelled with the classes and locations (with bounding boxes). The cataract video dataset is randomly split into a training set with 15 videos (13583 frames) and a testing set with 5 videos (3928 frames).

PSI-AVA dataset [21] includes eight radical prostatectomy surgeries, performed with the Da Vinci SI3000 Surgical System. It includes 11 phases, 21 steps, 16 actions and 7 instruments, and resolution is 1280×800. The phase and step annotations are provided by frames at 1 fps, while the actions and instruments are given by frames every 35 seconds of video. Thus, 73,618 keyframes are annotated with one phase-step pair, while 2238 keyframes have instruments and actions annotations. We split the dataset following work [21].

MultiBypass140 dataset [47] contains 140 Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) surgical videos from two medical centers, designed for surgical phase and step recognition. We selected the BernBypass70 subset (comprising 305,907 frames) from one center to conduct experiments and perform method comparisons on our model.

To demonstrate the generalizability of our method, we conduct additional experiments on the 50Salads [215] dataset. The 50Salads dataset contains 50 top-view videos of salad preparation activities performed by 25 different users in the same kitchen. The dataset provides detailed activity annotations at two levels: 3 high-level activities (preparing ingredients, making salad, and serving salad) and 17 low-level activities (e.g., cutting, peeling, mixing, dressing), which are directly analogous to the phase and step recognition tasks in our framework. The hierarchical nature of the dataset—high-level activities representing broader tasks, and low-level activities corresponding to specific actions—makes it an excellent fit for evaluating the multi-level semantic scene understanding capabilities of our proposed hierarchical context transformer.

5.4.2 Implementation Details

The proposed architecture is built with Pytorch and is trained on NVIDIA RTX A6000. MViTv2 [211] pretrained on Kinetics-400 [216] is adapted as the backbone. We use AdamW optimizer with a base learning rate of $1e-3$, and A cosine scheduler with a weight decay of 0.05. The model is totally trained for 60 epochs with a 10 epoch warm-up period. Data augmentation with 224×224 cropping, jittering and flipping is performed to enlarge the training dataset. The sequence length is 16 frames as [211]. The weights of the losses for phase, step, instrument and action tasks are set as 0.3, 0.2, 0.3 and 0.2 respectively. The batch size for both the baseline and the proposed method is consistent, set to 20. The temperature parameter in the ICL module is set to 0.07. The settings are the same for the two datasets. For getting the box-specific features, we pretrain DINO-5scale [213] on training data for our cataract datasets. For PSI-AVA dataset, due to the limited annotation of instruments, we conduct pretraining based on the EndoVis 2017 [152] and EndoVis 2018 dataset [153] following up [21], and then train on PSI-AVA. Bounding boxes are selected with a confidence threshold of 0.75, and corresponding 256-dimensional feature embeddings are extracted as box-specific features. To assess performance in the phase and step recognitions, we use the mean Average Precision (mAP) and Accuracy (Acc), the standard metrics in videos. For instrument and action detection, we use the mAP metric at an Intersection-over-Union threshold of 0.5 (mAP@0.5IoU).

5.4.3 Comparison with State-of-the-arts

To validate the performance of our model, we carry on comparisons between single-task models TMRNet [200], NETE [201] and SAHC [95] in phase and step recognition, as well as action detection ACTNet[163]. All three models TMRNet [200], NETE [201] and SAHC [95] focus on phase recognition, which are also employed to conduct step recognition for

Table 5.1: Performance comparison on the cataract for phase and step recognition. The best results are highlighted in bold. "Frames" denotes the number of frames used in models.

Method	Frames	Phase Recognition				Step Recognition			
		mAP	Acc	B-Acc	Re	mAP	Acc	B-Acc	Re
TMRNet [200]	30	0.9496	0.9358	0.8710	0.8710	0.9375	0.9230	0.8617	0.8617
NETE [201]	Video	0.9356	0.9312	0.9419	0.9419	0.8335	0.8921	0.7938	0.7938
SAHC [95]	Video	0.9288	0.9358	0.9363	0.9363	0.8982	0.9109	0.8083	0.8083
SlowFast [217]	16	0.9075	0.9153	0.9257	0.9257	0.8486	0.8983	0.7121	0.7121
TAPIR [212]	16	0.9286	0.9472	0.9448	0.9448	0.9191	0.9528	0.8970	0.8970
MViTv2 [211]	16	0.9494	**0.9557**	0.9284	0.9284	0.9289	0.9329	0.8829	0.8829
HCT	16	**0.9635**	0.9468	0.9454	0.9454	**0.9581**	**0.9556**	**0.9154**	**0.9154**

comparison in this chapter. To facilitate a comparison of phase and step tasks based on our dataset, we utilize existing methods to conduct separate recognition experiments for phase and step. TMRNet [200] introduces a memory bank to capture long-sequence temporal information, and even features from frames spanning even 30 seconds are utilized in our comparison. In our experiments, we utilize features from frames spanning even 30 seconds to obtain temporal dependencies. NETE [201] separates end-to-end training into an initial prediction step and a refinement stage. allowing the model to converge and optimize more effectively. Based on NETE [201], SAHC [95] is recently proposed to learn high-level segments from surgical videos, which are then used to correct ambiguities caused by low-level frame errors. Additionally, we perform multi-task performance comparisons based on various video recognition methods SlowFast [217], TAPIR[21] and MViTv2 [211] by adding the same multi-task losses and heads.

To comprehensively evaluate the effectiveness of our method, we introduced additional

Table 5.2: Performance comparison on the cataract for action and tool detection. The best results are highlighted in bold. "Frames" denotes the number of frames used in models.

Method	Frames	Action Detection		Tool Detection	
		mAP	Re	mAP	Re
ACTNet [163]	16	0.3970	0.8561	-	-
SlowFast [217]	16	0.4745	0.8703	0.8525	0.8141
TAPIR [212]	16	0.5251	0.8760	0.8541	0.8144
MViTv2 [211]	16	0.5563	0.8764	0.8543	0.8147
HCT	16	**0.5726**	**0.8776**	**0.8545**	**0.8149**

metrics such as Balanced Accuracy (B-Acc) and Recall (Re), as shown in Table 5.1, 5.2, 5.3 and 5.4. B-Acc adjusts for class imbalance by averaging the Recall of each class, providing a fairer performance evaluation. Recall measures the ability to correctly identify positive instances, highlighting its ability to capture relevant instances. The alignment of B-acc and Recall values is due to both metrics reflecting the correct recognition rate for each class for phase and step recognition.

5.4.3.1 Comparison with State-of-the-arts on Our Cataract

Table 5.1 and 5.2 present a performance comparison between HCT and the SOTA methods on our cataract dataset. The best results are highlighted in bold. Given that instrument detection is comparatively simpler and the current method has achieved high accuracy, our analysis is primarily focused on the other three tasks.

Using the cataract dataset, as shown in Table 5.1 and 5.2 TMRNet achieves the best performance among the three existing methods, indicating that longer video frames (30

Table 5.3: Performance comparison on the PSI-AVA for phase and step recognition. The best results are highlighted in bold. "Frames" denotes the number of frames used in models.

Method	Frames	Phase Recognition				Step Recognition			
		mAP	Acc	B-Acc	Re	mAP	Acc	B-Acc	Re
TMRNet [200]	30	0.5439	0.5734	0.4118	0.4118	0.4061	0.5357	0.3527	0.3527
NETE [201]	Video	0.4806	0.6403	0.4571	0.4571	0.2956	0.5116	0.3240	0.3240
SAHC [95]	Video	0.5522	0.6436	0.5183	0.5183	0.3798	0.5272	0.4197	0.4197
SlowFast [217]	16	0.4680	0.5463	0.3717	0.3717	0.3386	0.4713	0.2951	0.2951
TAPIR [212]	16	0.5824	0.5996	0.5516	0.5516	0.4561	0.5228	0.3860	0.3860
MViTv2 [211]	16	0.6189	0.6361	0.5512	0.5512	0.4837	0.5548	0.4452	0.4452
HCT	16	**0.6457**	**0.6603**	**0.5952**	**0.5952**	**0.4977**	**0.5766**	**0.4971**	**0.4971**

frames) significantly enhance phase and step recognition. However, compared to TMRNet, our method still achieves a 1.39% (96.35% vs. 94.96%) improvement in phase recognition and a 2.06% (95.81% vs. 93.75%) improvement in step tasks on mAP, even though our method only uses a sequence of 16 frames, while NETE [201] and SAHC [95] use whole video features in models. Furthermore, all three methods require at least two training stages on the training dataset, whereas our approach can achieve end-to-end recognition for both phase and step. Our model on action detection, compared with ACTNet, an anchor-context action detection network, also achieves a significant improvement in terms of mAP.

For the multi-task effect comparison, we adopt three popular SOTA methods in action recognition, encompassing both CNN-based and transformer-based approaches. By replacing the original action recognition heads of these methods with our multi-task heads, we facilitate a comparison of multi-task effects. To ensure fairness, neither the three methods nor our approach utilizes pre-trained models. The results demonstrate that our method outperforms

Table 5.4: Performance comparison on the PSI-AVA for action and tool detection. The best results are highlighted in bold. "Frames" denotes the number of frames used in models.

Method	Frames	Action Detection		Tool Detection	
		mAP	Re	mAP	Re
ACTNet [163]	16	0.2259	0.6854	-	-
SlowFast [217]	16	0.1857	0.5059	0.8029	0.7775
TAPIR [212]	16	0.2356	0.6881	0.8082	0.7783
MViTv2 [211]	16	0.2729	0.6929	0.8206	0.7896
HCT	16	**0.2820**	**0.6953**	**0.8217**	**0.7902**

others in every task. Specifically, in the tasks of phase, step and action, we exceed the performance of the best comparison method MViTv2 by 1.41% (96.35% vs. 94.94%), 2.92% (95.81% vs. 92.89%) and 1.63% (57.26% vs. 55.63%) respectively, highlighting the strong synergistic effect among these four tasks. Regarding the lower phase recognition accuracy on the cataract dataset compared to the baseline, we calculate the Balanced Accuracy (B-Acc) for both methods (MViTv2 92.84% vs. HCT 94.54%), as shown in Table 5.1 and 5.2. Our method shows a 0.89% lower Acc but a 1.7% higher B-Acc than MViTv2. This indicates that while MViTv2 performs better on more frequent categories, our approach achieves a more balanced performance across all categories. Notably, it excels in more challenging categories with fewer samples, making it more suitable for practical applications.

5.4.3.2 Comparison with State-of-the-arts on PSI-AVA

Regarding the PSI-AVA dataset, we conduct similar comparisons, as shown in Table 5.3 and 5.4. It's important to note that the PSI-AVA dataset, compared to our cataract dataset, has considerable differences in annotations for each task. Unlike our cataract dataset, where

Table 5.5: Performance comparison on phase recognition based on BernBypass70

Method	mAP	Acc	Re
CNN [177]	-	74.53 ± 13.34	45.69 ± 8.19
LSTM [218]	-	79.73 ± 13.75	56.19 ± 10.24
MT-LSTM [47]	-	80.69 ± 13.85	57.14 ± 13.38
TeCNO [64]	-	83.81 ± 13.55	62.81 ± 14.07
MTMS-TCN [25]	-	85.30 ± 13.19	67.41 ± 13.81
MViTv2 [211]	67.80	85.48 ± 8.90	67.83 ± 13.18
Ours	**68.59**	**86.52 ± 8.82**	**68.62 ± 12.39**

every frame is annotated with labels or bounding boxes for phase, step, instrument, and action at 1fps, the PSI-AVA dataset annotates only phase and step labels at 1fps, with instrument and action labels annotated at 35-second intervals. For a fair comparison, we follow the same training approach as in existing work [21]. It involves first training for phase and step recognition, followed by instrument and action detection using the best-trained step recognition model. This also results in the PSI-AVA dataset not benefiting from the mutual promotion of the four tasks during training, as observed in the cataract dataset training. For phase and step recognition tasks, the model is only optimized by the interaction between phase and step, and similarly for instrument and action detection. Nevertheless, our method still achieves the best performance in each task, further validating the effectiveness of our framework.

5.4.3.3 Comparison with State-of-the-arts on MultiBypass140

Following the five-fold cross-validation splits provided by the MultiBypass140 [47] dataset, we conducte experiments across various models, computing the average values and variances

Table 5.6: Performance comparison on step recognition based on BernBypass70

Method	mAP	Acc	Re
CNN [177]	-	58.92 ± 11.63	38.47 ± 7.39
LSTM [218]	-	64.99 ± 12.44	48.66 ± 11.12
MT-LSTM [47]	-	63.54 ± 13.92	48.49 ± 12.37
TeCNO [64]	-	67.54 ± 13.49	53.01 ± 11.74
MTMS-TCN [25]	-	67.54 ± 13.28	52.84 ± 10.44
MViTv2 [211]	45.93	72.51 ± 10.58	53.13 ± 12.34
Ours	**46.41**	**73.54 ± 10.61**	**54.79 ± 11.18**

across the five folds as shown in Table 5.5 and 5.6. To ensure a more comprehensive comparison, we have added the Recall metric on top of this chapter's evaluation metrics. The results indicate that our method consistently outperforms the baseline across all metrics, demonstrating the mutual enhancement between tasks in the multi-task framework. Moreover, when compared to the existing methods provided by [47], our approach outperforms them, particularly in the step recognition task.

5.4.3.4 Comparison with State-of-the-arts on 50Salads

In line with related works [224], the evaluation of the 50Salads dataset was conducted using five-fold cross-validation, with the average results reported. To better compare with existing methods, we additionally incorporate the segmental Edit distance score (Edit) on top of the original metrics. The Edit is commonly used in action segmentation tasks to evaluate how well predicted sequences align with the ground truth. A lower score indicates a closer match, reflecting higher recognition performance. As shown in Table 5.7, our method achieves the best performance in both high-level and low-level activity recognition tasks on this dataset.

Table 5.7: Performance Comparison on the Salad Dataset

Task	Method	mAP	Edit	Acc	Re
high-level activities	MViTv2	97.1	95.6	95.5	92.2
	Ours	**97.9**	**96.3**	**96.2**	**93.3**
low-level activities	MS-TCN [219]	-	67.9	80.7	-
	MS-TCN++ [220]	-	74.3	83.7	-
	UVAST [221]	-	83.9	79.5	-
	LGTNN [222]	-	82.0	86.1	-
	ASFormer [223]	-	79.6	85.6	-
	LTContext [224]	-	83.2	87.7	-
	MViTv2	87.0	83.7	87.5	81.6
	Ours	**87.7**	**84.3**	**88.4**	**83.4**

This allows us to demonstrate how well our hierarchical context transformer and inter-task contrastive learning generalize to natural scenes. We report the performance of our model in comparison to existing methods, providing insights into the model's effectiveness in broader applications.

5.4.4 Ablation Study

5.4.4.1 Contribution of Key Model Components

We conduct ablation experiments on two datasets to validate the contribution of key modules, including HRAM and ICL, with the results presented in Table 5.8. The models excluding the HRAM and ICL modules serve as our baseline models. It is built upon the MViTv2 framework, supplemented with multi-task loss and multi-task output heads. To minimize

Table 5.8: Ablation study on key Components based on the two datasets.

Modules		Phase	Step	Instrument	Action
HRAM	ICL				
Cataract dataset					
		0.9494	0.9289	0.8543	0.5563
✓		0.9523	0.9489	**0.8547**	0.5682
✓	✓	**0.9635**	**0.9581**	0.8545	**0.5726**
PSI-AVA dataset					
		0.6189	0.4837	0.8206	0.2729
✓		0.6251	0.4914	0.8213	0.2794
✓	✓	**0.6457**	**0.4977**	**0.8217**	**0.2820**

computational costs, all of our improvements are based on the final block. We compare the effects of adding HRAM to the baseline model, as well as the combination of HRAM and ICL modules. For the cataract dataset, adding HRAM improves the performance across all tasks. It enhances phase recognition, step recognition and action detection by 0.29% (95.23% vs. 94.94%), 2.00% (94.89% vs. 92.89%) and 1.19% (56.82% vs. 55.63%), respectively. This indicates a strong correlation and dependency among the four tasks within the cataract dataset. After further incorporating the ICL module, the performance of various tasks is further enhanced. This demonstrates that task-specific features are strengthened through intrinsic consistency extracted by the ICL module. The performance drop in instrument detection after incorporating the ICL module stems from limited task complementarity, as instrument detection, which depends on spatial precision, benefits less from shared temporal information across other tasks. Additionally, the multi-task learning framework is designed to prioritize overall task balance. Nonetheless, the results demonstrate that these effects are minimal in our model.

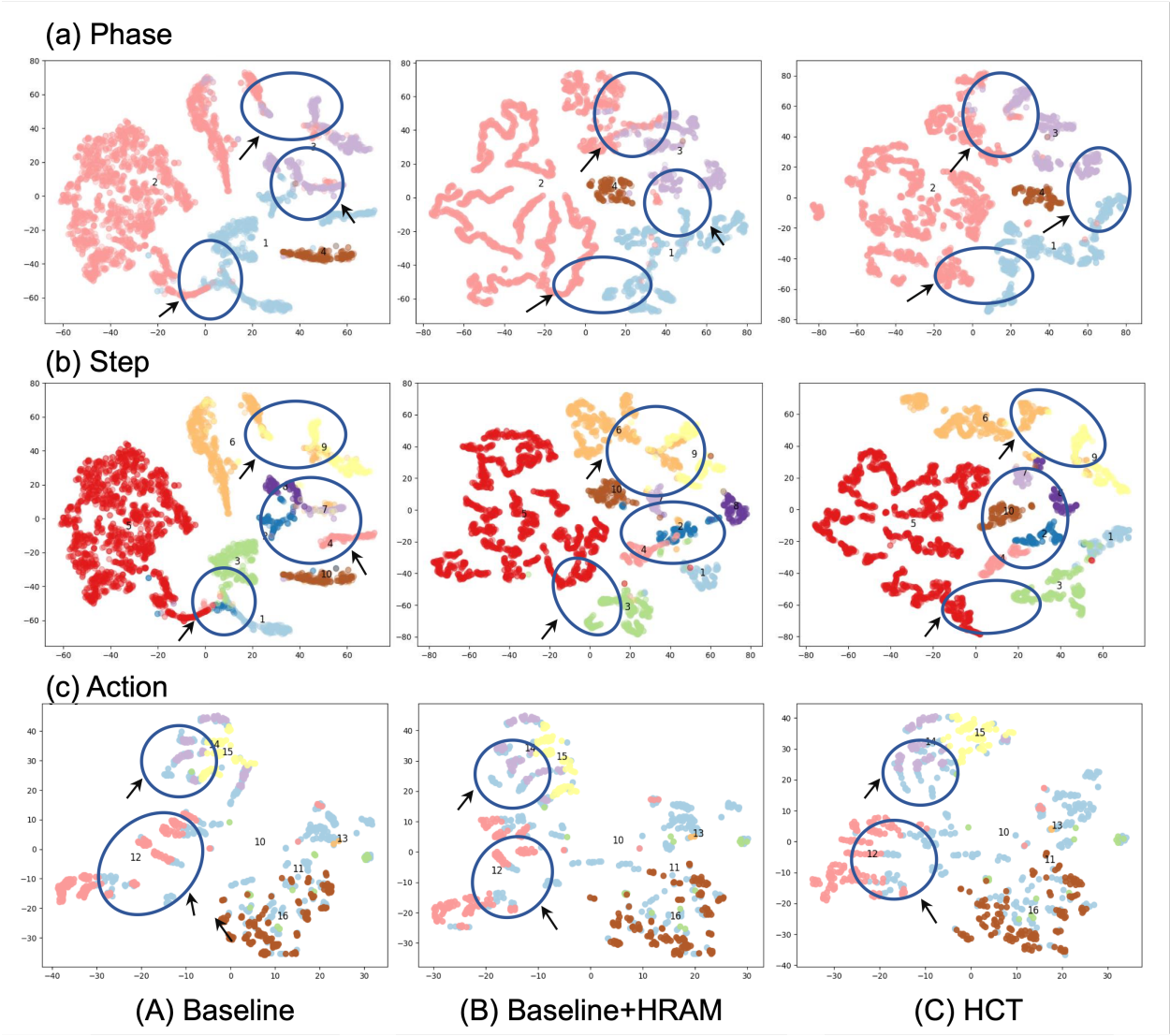


Figure 5.6: T-SNE Visualization results of (A) baseline, (B) baseline+HRAM and (C) HCT in phase, step and action recognition task. The blue circle indicates improvement in effectiveness. Due to space constraints, only the distribution of the 7 challenging action classes related to the instrument "Cannula" are shown in the figures, illustrating our method's enhanced performance on these difficult action classes.

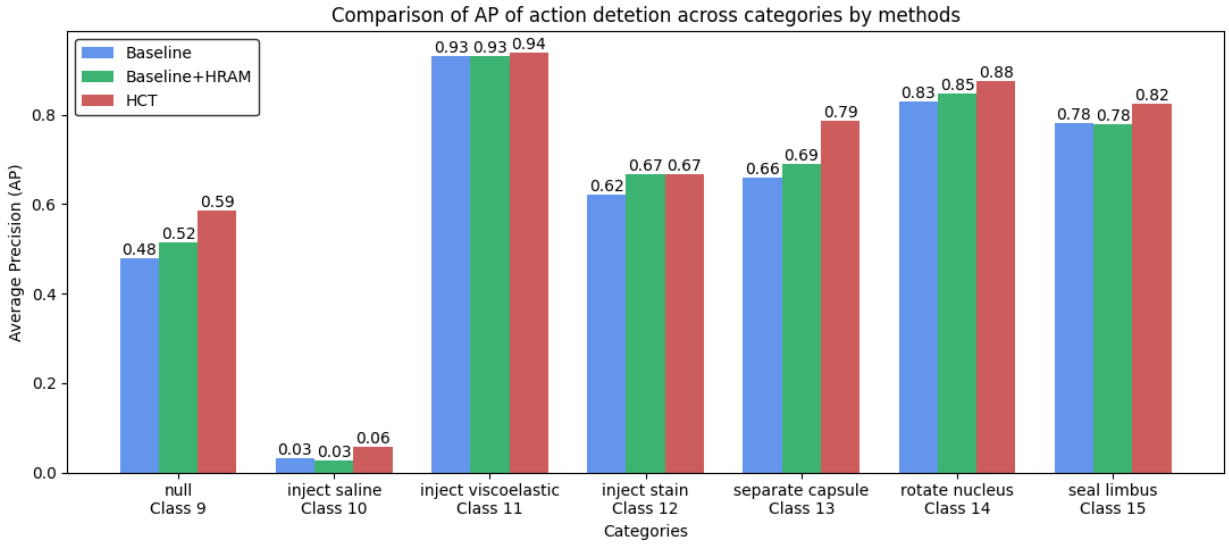


Figure 5.7: Comparision of Average Precision (AP) of action detection across the challenging classes related to the instrument "Cannula" in Figure 5.6, to quantitatively demonstrate enhanced effectiveness.

In parallel, for the PSI-AVA dataset, due to substantial differences in the distribution of labels for different tasks as described above, our experimental design follows [21] to implement the four tasks. This means that we have to divide the experiments into two parts: phase and step, and instrument and action. Consequently, the HRAM and ICL modules only achieve mutual enhancement between two tasks, rather than incorporating information from all four tasks. However, even under these constraints, the addition of the HRAM still results in superior performance compared to the baseline model. Further improvements are observed with the addition of the ICL module, with an improvement of even 2.68% (64.57% vs. 61.89%) in phase recognition against the baseline model.

t-SNE visualization results illustrating the contribution of our key model components are depicted in Figure 5.6. The blue circle in Figure 5.6 demonstrates that adding our key modules improves the distinguishability of classes and reduces inter-class blurring across phase, step, and action tasks. For the action detection task, due to space constraints, we

Table 5.9: Ablation study on HRAM.

Items				Phase	Step	Instrument	Action
P	S	I	A				
	✓	✓	✓	0.9153	0.9235	0.8544	0.5497
✓		✓	✓	0.9337	0.8739	0.8545	0.5561
✓	✓		✓	0.9045	0.8827	0.8534	0.5423
✓	✓	✓		0.9412	0.9046	0.8515	0.5471
✓	✓	✓	✓	**0.9523**	**0.9489**	**0.8547**	**0.5682**

focus on 7 classes generated by the instrumented cannula. These classes exhibit significant spatial similarity, posing classification challenges. The baseline model mixes these classes as shown in Figure 5.6 (c). Upon integrating our proposed module, the distinctiveness among these classes, highlighted by the blue circle, notably increases. To quantitatively validate this observation, we compare the Average Precision (AP) values of these seven classes across three models in Figure 5.7, aligning with their labeling in Figure 5.6, thus confirming the efficacy of our proposed module.

5.4.4.2 Contribution of HRAM

For a more detailed analysis of the HRAM module's role in enhancing interactions among various tasks, we conduct experimental analyses as presented in Table 5.9. "P", "S", "I", and "A" represent the tasks of phase, step, instrument, and action respectively. A checkmark beside a task indicates its participation in mutual enhancement, while its absence signifies exclusion. This approach allows us to analyze the impact of each task on the performance of a multi-task framework. Table 5.9 reveals that when the influence of "P" is excluded, the mAP for phase recognition is relatively lower (91.53%). When "S" is excluded, the performance

of step recognition is the poorest (only 87.39%). A similar pattern is observed in the action detection task. This suggests that the optimization of a task is significantly diminished when the task lacks participation in interactions with other tasks. This also underscores the importance of other tasks in optimizing or enhancing the specific task. Notably, when instrument features are missing, the performance of all phase, step, and action tasks is comparatively poor, highlighting the critical guiding role of instrument features for these three tasks. This aligns with our understanding of surgical videos, where certain specific instruments can directly indicate the current phase and step of a frame, and even some simple ongoing actions can be inferred. The last row of the table shows that when all tasks are incorporated into our framework and mutually enhance each other, we achieve optimal performance for all four tasks with even mAP of 95.23% and 94.89% for phase and step.

5.4.4.3 Contribution of ICL

To verify the effectiveness of ICL, we carry on detailed ablation experiments based on the cataract dataset. Since ICL optimizes the model by calculating losses for pairwise task features, we design ablation experiments in two aspects (separated by double solid lines in Table 5.10) to assess the performance differences of each task within ICL. The table lists six pairs of our four tasks, namely "PS", "PI", "PA", "SI", "SA", and "IA". On one hand, similar to HRAM, we calculate ICL by excluding the influence of one task. Table 5.10 (above the double solid lines) shows the model's performance when calculating ICL pairwise among the remaining three tasks after excluding one task ("P", "S", "I", or "A"). We observe that relatively lower results are yielded for a task when it is excluded by ICL. In Table 5.10, excluding "P" leads to the poorest performance in the phase recognition (only 92.08%). Excluding "S" results in the second-worst performance in step recognition. The absence of "I" leads to lower performance in phase and step recognition, as well as instrument detection. The mAP for step recognition even drops to only 87.59%. Similarly, excluding

Table 5.10: Ablation study on ICL. The best results are highlighted in bold, while the next best results are in underline.

Items						Phase	Step	Instrument	Action
PS	PI	PA	SI	SA	IA				
			✓	✓	✓	0.9208	0.9145	0.8539	0.5472
	✓	✓			✓	**0.9672**	0.9113	<u>0.8547</u>	<u>0.5562</u>
✓		✓		✓		0.9372	0.8759	0.8532	0.5481
✓	✓		✓			0.9446	0.9367	0.8546	0.5398
✓	✓	✓	✓	✓	✓	0.9353	0.9240	**0.8554**	0.5531
		✓	✓			0.9532	0.9124	0.8543	0.5535
	✓			✓		0.9580	<u>0.9452</u>	0.8544	0.5478
✓					✓	<u>0.9635</u>	**0.9581**	0.8545	**0.5726**

"A" reduces the performance of action detection. This confirms the importance of each task in the functioning of ICL.

However, due to inherent differences between tasks, the optimal results are not able to be yielded when all four task pairs are fed into ICL, as shown in Table 5.10. In our tasks, phase and step have similar optimization objectives and can extract the strongest consistencies for the same positive sample, as do action and instrument. Our second set of experiments, as illustrated in Table 5.10 (below the double solid lines), confirms this. When we optimize the model by combining the pairs of "PS" and "IA", the model achieves optimal performance in each task, further improving upon the addition of HRAM.

Table 5.11: Performance comparison on new baseline model with the cataract dataset for four tasks

Method	Phase Recognition				Step recognition			
	mAP	Acc	B-Acc	Re	mAP	Acc	B-Acc	Re
MViTv2*	0.9489	0.9501	0.9129	0.9129	0.9191	0.9528	0.8970	0.8970
MViTv2	0.9494	**0.9557**	0.9284	0.9284	0.9289	0.9329	0.8829	0.8829
HCT	**0.9635**	0.9468	**0.9454**	**0.9454**	**0.9581**	**0.9556**	**0.9154**	**0.9154**
Method	Action detection		Tool detection	
	mAP	Re	mAP	Re
MViTv2*	0.5441	0.8763	0.8537	0.8148
MViTv2	0.5563	0.8764	0.8543	0.8147
HCT	**0.5726**	**0.8776**	**0.8545**	**0.8149**

5.4.4.4 Performance Comparison on New Baseline Model

The original MViTv2 model consists of 16 blocks. According to the granularity of the tasks, we use the output of the 14th block for phase recognition, the output of the 15th block for step recognition, and the output of the final block for instrument and action detection. Thus, we form a new baseline model. As shown in Tables 5.11 and 5.12, MViTv2* indicates the results of the new baseline model. The new baseline model performs similarly to the original baseline, and our method consistently outperforms both baseline models.

Table 5.12: Performance comparison on new baseline model with the PSI-AVA for four tasks

Method	Phase Recognition				Step recognition			
	mAP	Acc	B-Acc	Re	mAP	Acc	B-Acc	Re
MViTv2*	0.5942	0.5915	0.5042	0.5042	0.4739	0.5141	0.3360	0.3360
MViTv2	0.6189	0.6361	0.5512	0.5512	0.4837	0.5548	0.4452	0.4452
HCT	**0.6457**	**0.6603**	**0.5952**	**0.5952**	**0.4977**	**0.5766**	**0.4971**	**0.4971**
Method	Action detection				Tool detection			
	mAP		Re		mAP		Re	
MViTv2*	0.2754		**0.7086**		0.8206		**0.7909**	
MViTv2	0.2729		0.6929		0.8206		0.7896	
HCT	**0.2820**		0.6953		**0.8217**		0.7902	

5.4.5 Effectiveness of HCT+

Utilizing the Kinetics-400 [216] pretrained model provided by MViTv2 [211], we conduct experiments to validate the effectiveness of the ST-Ada, as presented in Table 5.13. The peak memory is measured when the batch size is 16. "Full-tuning" indicates fine-tuning of the entire model. As a reference baseline method, we perform full fine-tuning based on the pretrained model of the MViTv2 framework, which gets the poorest performance. When fully fine-tuning our proposed model, the performance on the four tasks reaches the optimum, but the "param" is also the highest at 57.2M, and memory consumption is at its peak. If we freeze the parameters of the first 15 transformer blocks of our proposed model and only optimize the last improved block, denoted as "w/o ST_Ada", the number of parameters to be optimized decreases by 57.6%. However, the performance on each task also decreases, with a notable drop of 4.60% (91.75% vs. 96.35%) in phase recognition.

Table 5.13: Performance on the cataract dataset for HCT+. "Param" denotes the number of parameters and "PM" denotes peak memory usage.

Method	Params(M)	PM(MB)	Phase	Step	Instrument	Action
MViTv2(Full-tuning)	35.7	34382	0.9494	0.9289	0.8543	0.5563
HCT(Full-tuning)	57.2 (100%)	42850	**0.9635**	**0.9581**	0.8545	**0.5726**
HCT(w/o ST_Ada)	24.3 (42.4%)	32816	0.9175	0.8967	0.8541	0.5412
HCT(w/ S_Ada)	24.6 (+1.2%)	32824	0.9381	0.9301	0.8539	0.5613
HCT(w/ T_Ada)	26.6 (+9.5%)	35610	0.9345	0.9271	<u>0.8545</u>	0.5537
HCT+(w/ ST_Ada)	26.9 (+10.7%)	35616	<u>0.9542</u>	<u>0.9437</u>	**0.8546**	<u>0.5672</u>

Nevertheless, with the addition of the spatial adapter (w/S_Ada), the model's performance experiences a widespread improvement, with only a marginal increase of 1.2% in parameters based on the "w/o ST_Ada" model. Additionally, during training, peak memory decreases by nearly 10G compared to our full-tuning model, significantly reducing computational costs. Similarly, when only the temporal adapter is added (w/T_Ada), with a minimal increase of 9.5% in parameters, the performance on each task is better than without any adapters. This indicates the excellent transferability of spatial and temporal adapters. Upon combining spatial and temporal adapters (HCT+), our model's performance further improves with only a 10.7% increase in parameters. The accuracy of phase and step recognition even surpasses that of "w/o ST_Ada" by 3.67% (95.42% vs. 91.75%) and 4.70% (94.37% vs. 89.67%), respectively. Here is just a preliminary exploration of the application of transfer learning in our task. Its feasibility further makes it possible to apply our multi-level semantic scene understanding task facts to surgical scenarios.

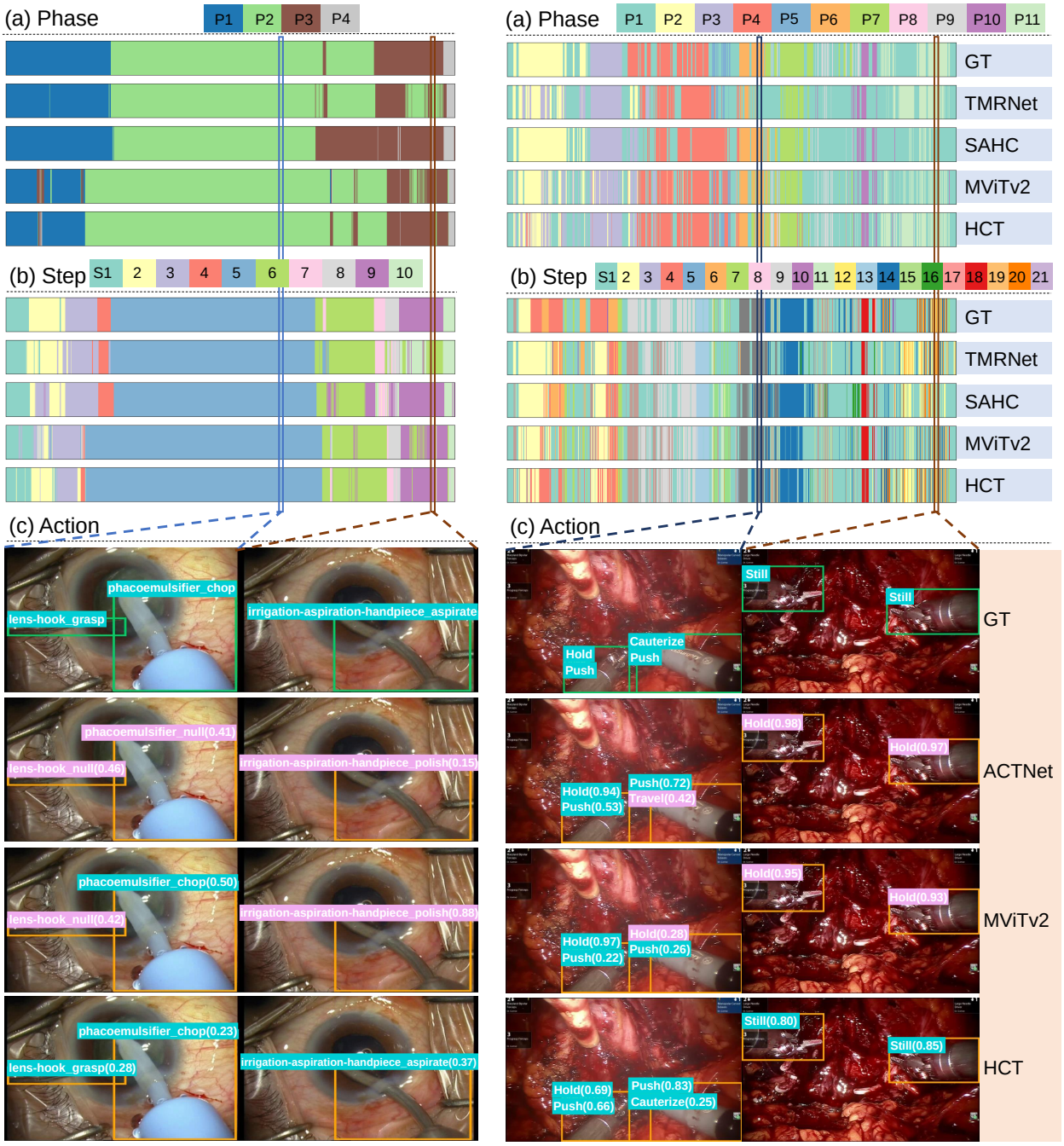


Figure 5.8: Multi-level visualisation based on (A) Video66 from cataract dataset and (B) CASE015 from PSI-AVA dataset. (a) Qualitative comparisons for phase recognition. (b) Qualitative comparisons for step recognition. (c) Qualitative comparisons for action detection. We show the Ground-Truth (GT) (sky blue), right predictions (sky blue) and wrong predictions (light purple).

5.4.6 Result Visualisation and Analysis

To illustrate the effectiveness of HCT more clearly, we list visualizations of prediction results for the two datasets in Figure 5.8, which presents the multi-level semantic understanding of surgical videos spanning coarse to fine-grained tasks. Limited by the page, we only list the partial results with high accuracy, TMRNet [200], SAHC [95] and MViTv2 [211] for the phase and step recognition, while ACTNet [163] and MViTv2 [211] for action detection. In Figure 5.8 (A), for the cataract dataset, different colours for phases and steps represent different classes. The first row provides the Ground Truth (GT) for reference. By examining the timelines of predicted phases and steps, it is evident that our approach exhibits excellent performance even for some challenging classes. For instance, in the step recognition, the S2, highlighted in yellow, constitutes a small proportion of frames in the entire video. Other methods misidentify this category, while our method can accurately predict it. In the action detection task, we visualize the predicted positions of actions and their corresponding classes for each method, using bounding boxes and text in the top-left corner of the boxes. The ground truth action positions are outlined in green in the first row, while the predicted positions by each method are shown in yellow. The GT action classes and correctly predicted classes are highlighted in sky blue, whereas incorrectly predicted classes are displayed in light purple. We also present the corresponding scores for predicted action. Detecting fine-grained actions is more challenging than phase and step recognition, as spatial features for some actions are highly similar, which is also illustrated by the performance of comparison methods. For example, in the left column of Figure 5.8 (A), cases such as "lens hook grasp" and "lens hook hold", appear visually similar and can be easily misjudged based on appearance alone. Correct judgment necessitates the integration of contextual information, specifically the temporal features of instruments in consecutive frames. Notably, our method performs accurately when other approaches struggle to predict the current action, underscoring the effectiveness of our method in learning temporal dynamics. To further demonstrate pro-

posed method's performance in accurately identifying positive instances compared to other methods, we computed the Recall metrics for all models, as detailed in Table 5.1, 5.2, 5.3 and 5.4. The results demonstrate that our method achieves the best Re performance.

For the PSI-AVA dataset, as depicted in Figure 5.8 (B), the comparison of phase and step recognition reveals that our method produces prediction results very close to the GT, particularly in regions with dense class distributions, which are also challenging to identify. The action detection task further demonstrates the excellent performance of our method in detecting multiple classes of concurrent actions.

5.4.6.1 Result Visualisation with More Cases

With more test cases, visual comparisons (Figure 5.9 and (Figure 5.10) are provided to show the improvement in surgical phase and step recognition qualitatively on the cataract and PSI-AVA dataset. In the figure, we show the results obtained from two challenging videos with different durations. We compare the proposed method with the six state-of-the-art methods including TMRNet [200], NETE [201], SAHC [95], slowfast [217], TAPIR [212] and MViTv2 [211]. From this figure, one can see that for both cases for the two datasets, the proposed method achieves more accurate predictions than the SOTA methods with more reliable predictions.

To illustrate the effectiveness of HCT more clearly, we present visualizations of prediction results for the cataract dataset and PSI-AVA dataset in Figure 5.11 and Figure 5.12, respectively. To demonstrate a multi-granularity understanding of surgical videos, our visualizations present, from top to bottom, the coarse-grained phase recognition results, step recognition results, and finer-grained action detection results for the same test video. To facilitate a clear and effective comparison, we select state-of-the-art and recently developed methods for visual comparison. The comparison methods for phase and step are same with

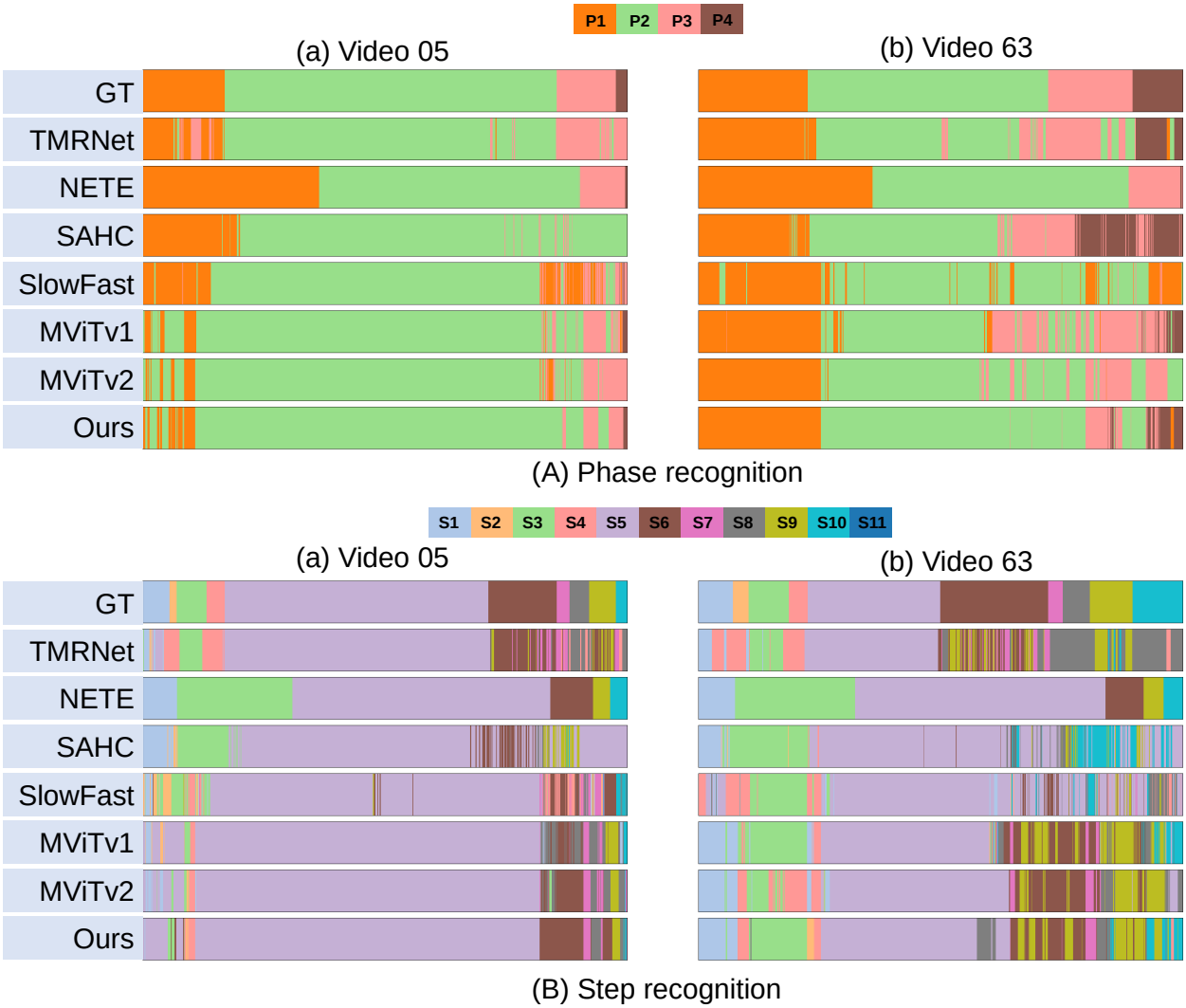


Figure 5.9: Qualitative comparison with our method and the state-of-the-art methods on two test video of cataract dataset for (A) phase recognition and (B) phase recognition. The color-coded ribbons illustrate the phases (P1-P4) and steps (S1-S11) during two videos with different durations. Horizontal axis represents the time progression. GT: Ground Truth.

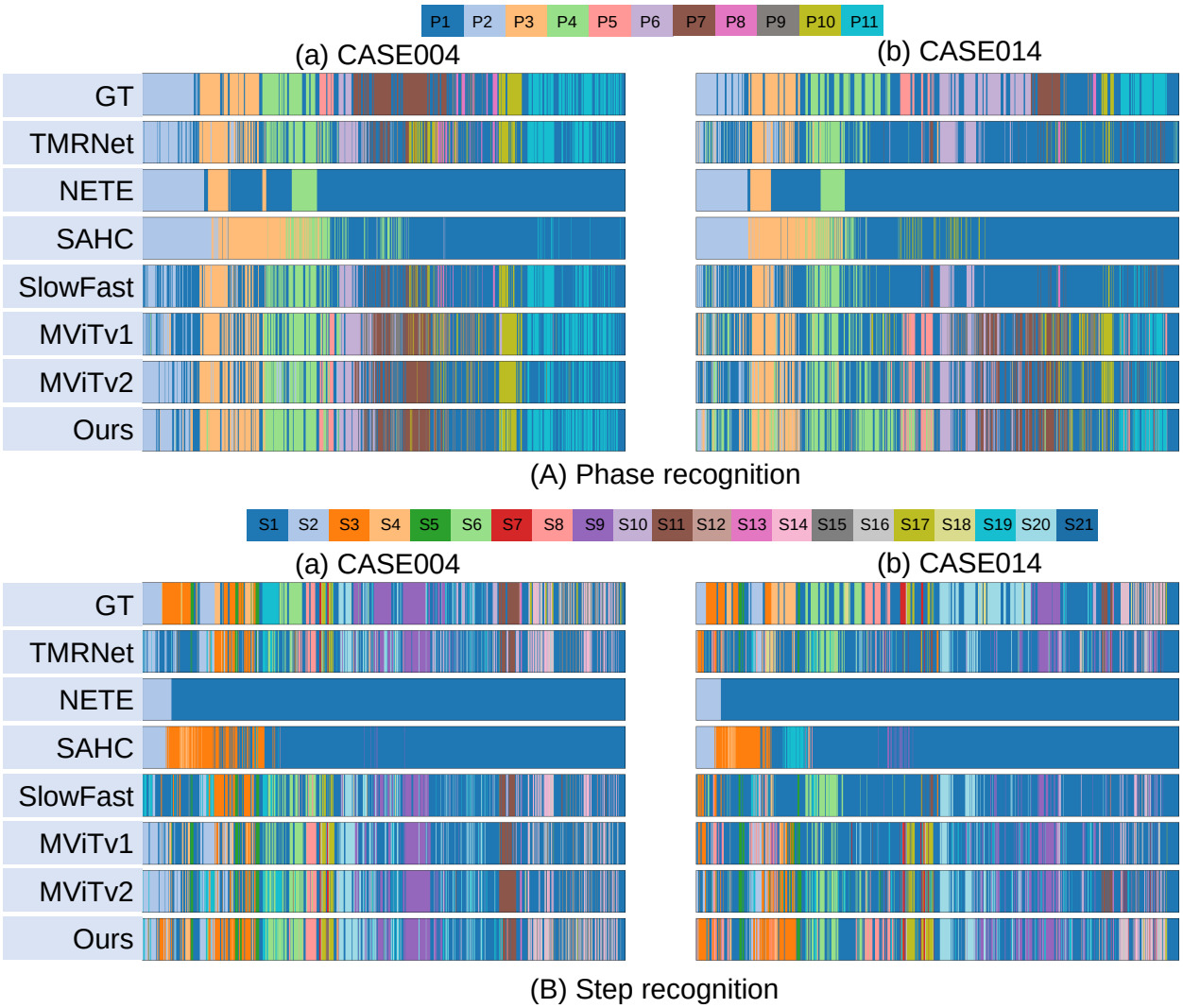


Figure 5.10: Qualitative comparison with our method and the state-of-the-art methods on two test video of PSI-AVA dataset for (A) phase recognition and (B) phase recognition. The color-coded ribbons illustrate the phases (P1-P11) and steps (S1-S21) during two videos with different durations. Horizontal axis represents the time progression. GT: Ground Truth.

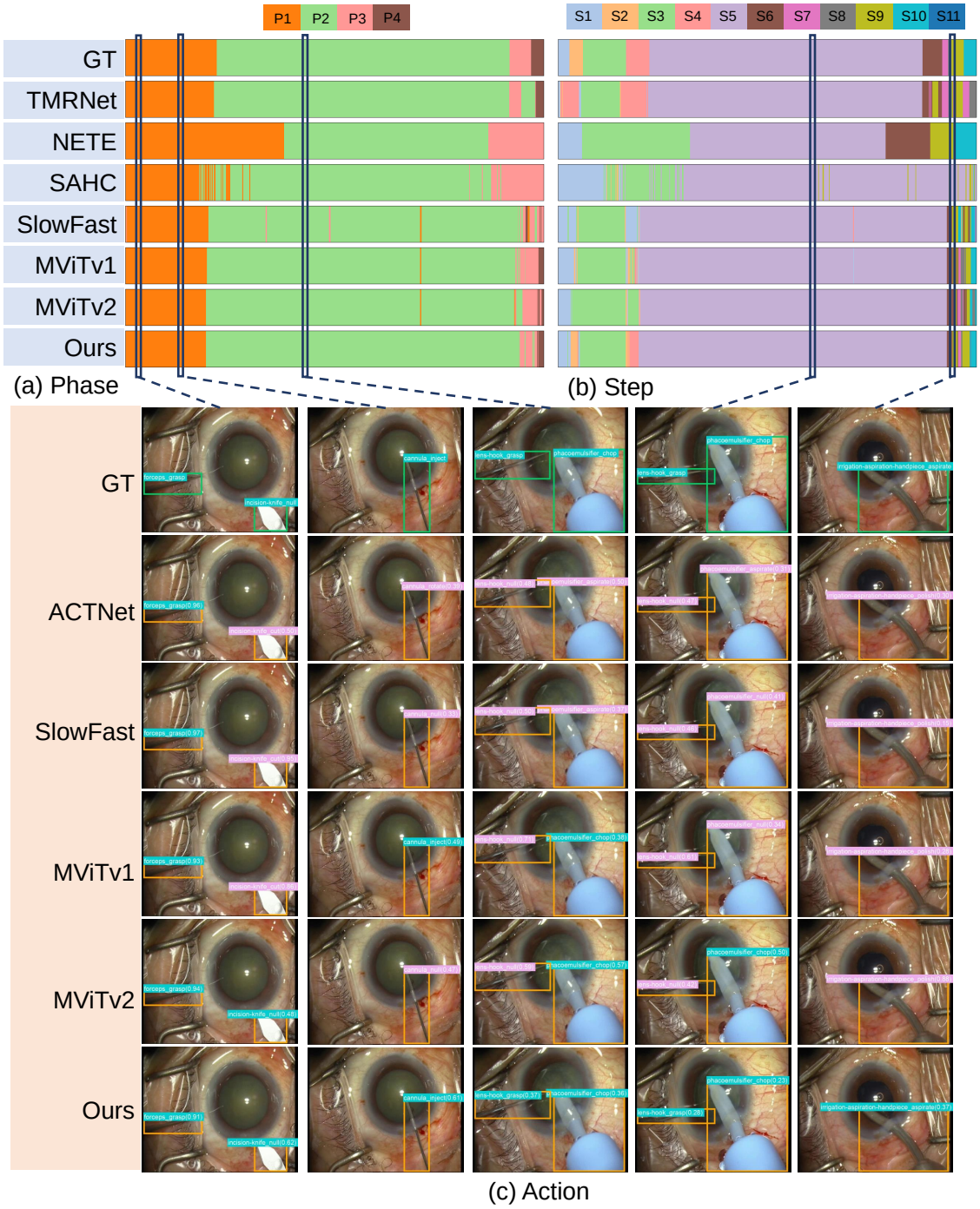


Figure 5.11: Multi-granularity visualisation based on Video30 from cataract dataset. (a) qualitative comparisons with the other six other methods for phase recognition. (b) qualitative comparisons with the other six other methods for step recognition. (c) qualitative comparisons with the other four methods for action detection. We show the ground-truth (sky blue), right predictions (sky blue) and wrong predictions (light purple). GT is the Ground Truth.

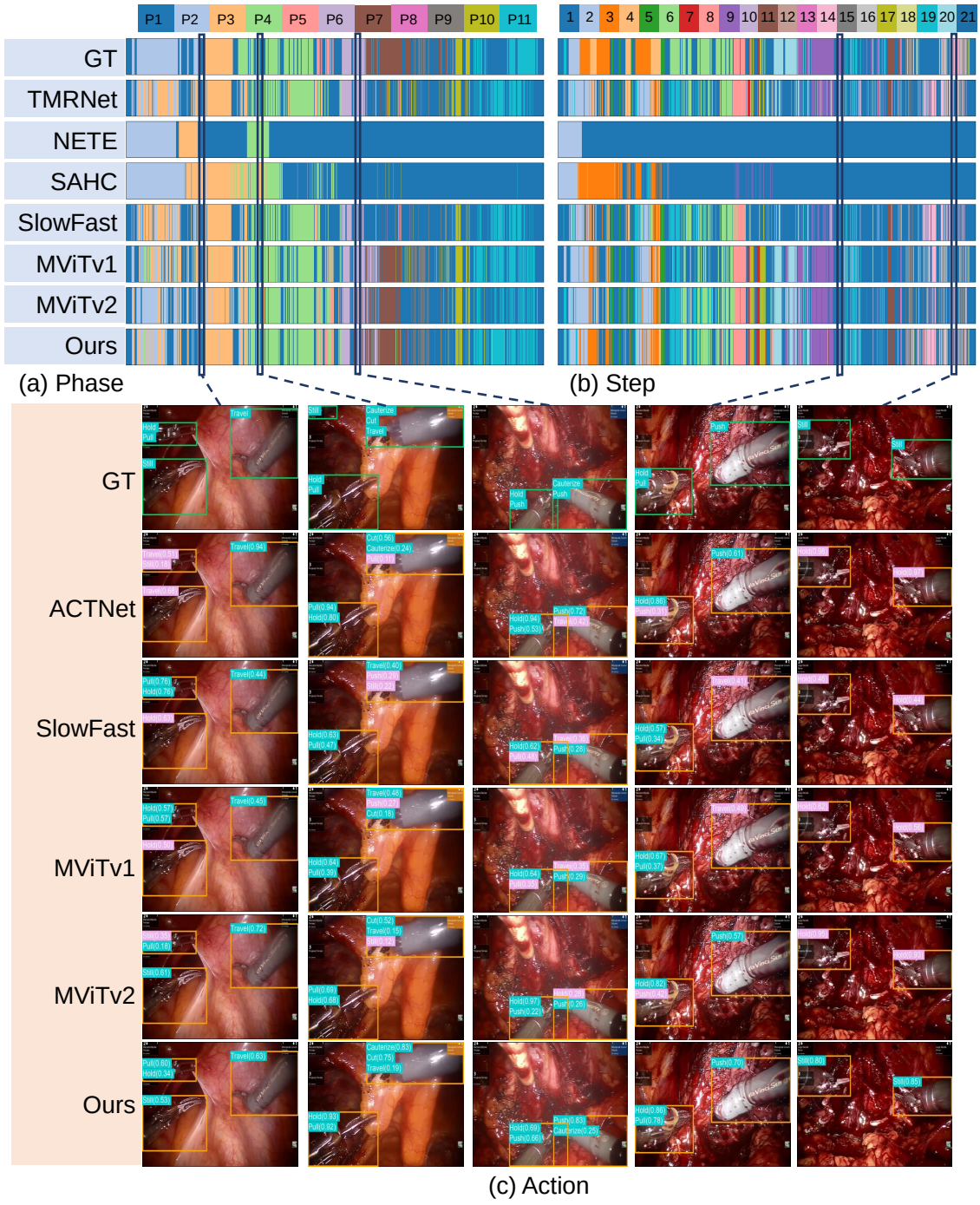


Figure 5.12: Multi-granularity visualisation based on CASE015 from PSI-AVA dataset. (a) qualitative comparisons with the other six methods for phase recognition. (b) qualitative comparisons with the other six methods for step recognition. (c) qualitative comparisons the other four methods for action detection.We show the ground-truth (sky blue), right predictions (sky blue) and wrong predictions (light purple). GT is the Ground Truth.

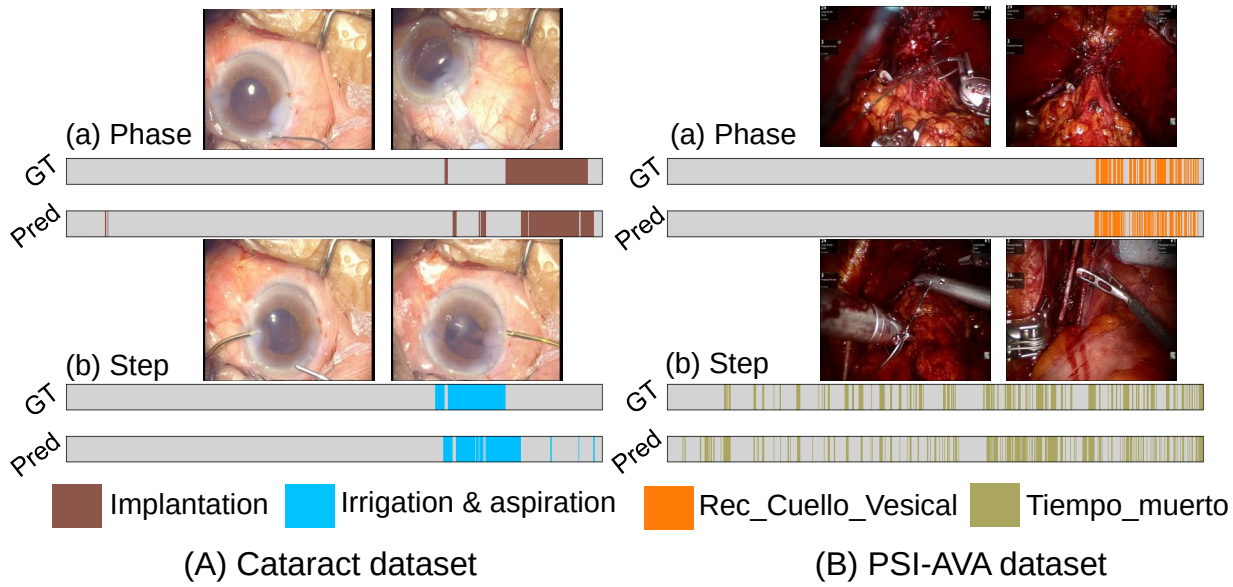


Figure 5.13: Color-coded ribbon visualization of phase and step recognition for some challenging classes from the proposed method (Pred) and the Ground Truth (GT) in two complete surgical videos. In each case, we show the classes with different colors and no class appearance with blank. Horizontal axes indicate the time progression.

Figure 5.9 and Figure 5.10, while slowfast [217], TAPIR [212], ACTNet [163] and MViTv2 [211] are selected for action detection. From the figure, it is evident that in the same case, not only does our method closely align with the ground truth in the prediction distribution of phases and steps, but it also excels in action detection. While other methods may inaccurately predict the portion or even the entirety of actions in the current frame, our method consistently predicts the occurring actions accurately. This is especially notable in the case of multi-class action detection, as shown in Figure 5.12.

5.4.7 Challenge Cases Analysis

We demonstrate the recognition performance of our model across phases and steps for challenging classes, as shown in Figure 5.13. The classes are represented in different colors, with

blank spaces indicating no class appearance. For the cataract dataset, as depicted in the video frame in Figure 5.13 (A).a, it is difficult to extract effective spatial information for discrimination. This is because the “Cannula” in the left figure is at the edge of the field of view, and the "Implant injector" in the right figure is transparent. In Figure 5.13 (A).b, the incomplete display of the "Handpiece" head instrument can lead to confusion with similar instruments, such as “Cannula”, posing a classification challenge.

In the PSI-AVA dataset, both the phase and step class distribution shown in Figure 5.13 B are scattered, making it challenging to differentiate classes using temporal information. In Figure 5.13 (B).a, the confusion between instruments and tissue in the video frame complicates the distinction of operations. In Figure 5.13 (B).b, the step "Tiempo_muerto" indicates the absence of a step, which is widely dispersed and scattered in the whole video, making it more difficult to identify null operations compared to other operations. Despite these challenges, our method predicts a distribution very similar to the GT. Combined with the results shown in Figure 5.7, this demonstrates the robustness and effectiveness of our method for all tasks.

5.5 Summary

In this chapter, we present a novel Hierarchical Context Transformer (HCT) network to access MSSU by achieving recognition of phases and steps, and detection of actions and instruments. Specifically, we develop a Hierarchical Relation Aggregation Module (HRAM) to dig out the intrinsic relationships between different tasks and further boost task-wise representation learning with the proposed Inter-task Contrastive Learning (ICL). For the costs of model training and storage on the transformer models, we introduce spatial and temporal adapters to equip our HCT for HCT+ with spatial-temporal reasoning capabil-

ity, which enables the HCT+ to achieve comparable performance with the full fine-tuned model at a small parameter cost. Comprehensive experimental results on two surgical video datasets demonstrate the effectiveness of our proposed model and the contribution of each key component.

Chapter Six

Deconfounded Multi-level Surgical Video Scene Understanding with Causal Intervention

This chapter is based on the following publications:

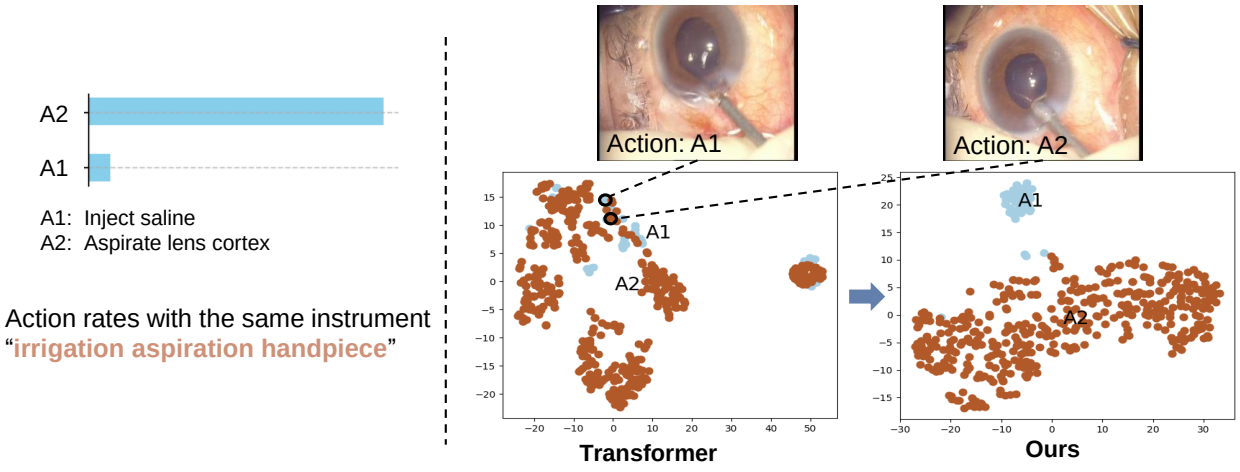
Deconfounded Surgical Video Representation Learning with Causal Intervention, which is under review in IEEE Transactions on Medical Imaging (2025).

Chapter 5 proposed the Hierarchical Context Transformer (HCT) framework, demonstrating that modeling the inter-task relationships between phases, steps, and actions significantly improves multi-level recognition. While HCT captures statistical co-occurrences, it, like most deep learning models, is susceptible to learning spurious correlations from dataset biases. For example, a model might spuriously associate an instrument with a phase simply because they frequently appear together, rather than learning a true causal link. This reliance on statistical shortcuts limits model robustness and generalization. This chapter tackles this fundamental problem by introducing a causality-inspired framework to deconfound video representations.

6.1 Introduction

While the hierarchical frameworks for CAS explored in the previous chapter are vital for enhancing healthcare and training [12, 163, 166], a critical limitation of standard correlation-based models is their susceptibility to dataset biases and spurious correlations [86, 108]. To achieve a truly robust and reliable understanding of surgical scenes [45, 166], CAS systems must move beyond simple correlations and interpret activities across multiple temporal granularities, ranging from phases, steps, to fine-grained actions [21, 189]. Thus, deconfounded multi-granularity recognition [45, 166] is essential for enabling CAS systems to offer richer, more reliable context-aware assistance to the surgical team.

To achieve multi-granularity recognition, it is crucial to develop a unified spatio-temporal representation capable of capturing both high-level phase information and fine-grained action details. Such representations are necessary to bridge the gap between phase, step, and action recognition, providing a robust foundation for accurate and adaptive surgical video understanding. Although existing methods have achieved excellent performance on specific



(a) Action occurrence rates **(b) T-SNE results comparison**

Figure 6.1: Examples about the spurious correlation in surgical video understanding. We show the example from the action detection task in Cataract_ours. (a) shows the action concurrent rates with the instrument "irrigation aspiration handpiece". (b) shows the comparison of the t-SNE results for the transformer model and ours. In this case, the transformer learns the spurious correlation and struggles to separate two actions. Thanks to causal inference, ours achieves clearer clustering.

tasks at different levels [64, 95, 100, 200]. They rely heavily upon fitting the data distributions and tend to capture the spurious correlations from features and predictions, and thus fail to learn the essential causal relations behind the knowledge that have a good generalization and cognitive abilities [86]. Recently, causal reasoning has attracted increasing attention in a myriad of high-impact domains of computer vision, such as long-tail classification [225, 226], image semantic segmentation [227], emotion recognition[228], visual dialog [229], image captioning [120] and visual question answering [117, 230, 231]. These studies reveal an underlying problem: neural networks, which are statistics-based and "lazy", often take shortcuts by focusing on contextual cues rather than the core features specific to each instance. When the training data is limited (e.g., few-shot learning) or deficient (e.g., long-tail distribution), the problem is more severe and obvious. In surgical video understanding,

we encounter a similar common issue where models frequently learn correlations by focusing on basic visual information rather than causal relationships. Traditional statistical learning methods capture such correlations effectively in ideal independent and identically distributed settings but falter when these conditions are not met [86]. For example, in Figure 6.1 (a)(b), in surgical action detection, both actions A1 (inject saline) and A2 (aspirate lens cortex) are performed with the same instrument "irrigation aspiration handpiece", with action A2 occurring much more frequently than A1 (51.7% v.s. 4.1%). In (b), the t-SNE visualization shows that the transformer model struggles to separate actions A1 and A2, leading to overlaps. Two original images from two actions with closely positioned feature points are also shown. The static visual features of the two frames appear highly similar. Therefore, when the "irrigation aspiration handpiece" appears, the transformer is more likely to classify the images as A2 rather than A1. However, with the introduction of causal inference, our method can correctly identify the images and achieve clearer clustering.

When the networks find the simple visual information is enough to recognize most of the activities, which represent phases, steps or actions, they often do not focus on the core representations specific to each instance. This example illustrates the significance of causal reasoning in removing biases and understanding the true relationships within surgical videos. In this work, we revisit surgical video understanding from a causal perspective, to uncover the fundamental relationship between features and predictions.

To this end, we propose a unified model designed to enable different levels of understanding in surgical videos, ranging from coarse-grained phase recognition and step recognition to fine-grained action detection. Firstly, our framework introduces a subject-guided representation module to extract key spatio-temporal features. In surgical videos, regardless of the level of activities, instruments (subjects) are the primary drivers of procedural activities. Emphasizing the key elements of the surgery offers a more efficient, compact representation. Our intuition is also supported biologically that we humans perceive the world by concen-

trating our focus on key regions in the scene [232, 233]. Additionally, we incorporate a causal inference module to disentangle true causal relationships from confounding factors, ensuring robust predictions. Unlike traditional video analysis approaches, our method emphasizes activity-relevant features while de-emphasizing noisy or irrelevant data, achieving improved effectiveness and generalizability.

In summary, our contributions are five-fold:

- We discover a long-overlooked issue that visual bias in surgical datasets leads existing models to learn spurious correlations (e.g., between static instrument appearance and common actions) rather than essential causal relations.
- We present a general Deconfounded Surgical Video Representation Learning (DSVRL) architecture, providing a unified, causality-inspired solution for multi-granularity surgical scene understanding.
- We first propose that a robust representation must be subject-centric, as instruments (subjects) are the primary drivers of surgical activity (the desired $X \rightarrow S \rightarrow Y$ causal path). We develop a Subject-Centric Module (SCM) to explicitly model these critical instrument-tissue interactions.
- We then address the key challenge introduced by this subject-centric focus: the SCM's representation is highly vulnerable to confounding bias from static instrument appearance (the spurious $X \leftarrow C \rightarrow S \rightarrow Y$ path). We therefore develop a surgical Activity Causal Intervention Module (ACIM) that applies backdoor adjustment directly to the SCM's representation, disentangling the true causal effect of the action from the confounding appearance bias.
- We conduct experiments with the proposed DSVRL network on four datasets across three tasks, demonstrating its state-of-the-art (SOTA) performance. We validate that

the SCM and ACIM work synergistically to improve model robustness, and cross-dataset experiments confirm the model's superior generalization.

6.2 Related work

6.2.1 Surgical Scene Understanding

Recent studies in surgical scene understanding generally fall into two main areas. The first focuses on phase or step recognition within untrimmed surgical videos, which has become a highly active research topic. For example, TMRNet [200] leverages a memory bank to capture long-term temporal features and improve phase recognition accuracy. Inspired by the MS-TCN [234] model's success in human action segmentation, Yi et al. [201] proposed a non-end-to-end training strategy based on MS-TCN to enhance surgical phase segmentation. TeCNO [64] also adaptes the MS-TCN [234] to online surgery scenarios via causal and dilated convolutions. With the rise of transformers for sequential data processing, Pan et al. [202] integrated Swin transformer and LSTM to extract multi-scale visual and temporal features for phase recognition in surgical videos. Czempiel et al. introduced OperA [92], a transformer-based model with an attention regularization loss to select high-quality frames for phase identification, while Ding et al. [95] developed SAHC to improve phase recognition by focusing on segment-level semantics. Trans-SVNet [235] aims to fuse spatial and temporal features by developing a small transformer-based fusion head. In another approach, Yue et al. [203] combined frame-level and phase-level temporal context to analyze surgical workflows comprehensively.

The second area addresses fine-grained action recognition. Bawa et al. [97] introduced the SARAS endoscopic minimally invasive surgery dataset, which addresses the challenge of

surgeon action detection and includes baseline results using open-source techniques such as Faster R-CNN [178] and YOLO [99]. Nwoye et al. [100] proposed representing actions as triplets (instrument, verb, target) in a laparoscopic dataset, employing a trainable 3D interaction space to model the relationships between these triplet elements. They later developed the Rendezvous (RDV) model [44], which leverages class activation-guided and mixed attentions to detect each component of the triplet. After adding the location information of the instrument and target, Lin et al. [102] presented a novel QDNet for the instrument-tissue interaction quintuple detection task in cataract videos. They further develop the ITIDNet [103] model to capture relations between intra- and inter-frame of instruments and tissues. To enhance action detection accuracy, Hao et al. [163] presented ACTNet, an anchor-based model that predicts action locations, classes, and confidence levels.

Additionally, several studies have sought to improve surgical video understanding by incorporating instrument detection with multi-task learning. For instance, Jin et al. [105] showed consistent improvement in both phase recognition and tool recognition tasks with multi-task learning. Tao et al. [236] introduced a multi-task transformer model for surgical phase recognition and tool detection, leveraging video-level semantic structure through a latent space-constrained approach. Yuan et al. [237] proposed an instrument interaction-aware network for anticipating surgical workflow, leveraging spatial context and long-term temporal dependencies. Hao et al. [163] presented HCT to explore the relations across the different level tasks. The core of these methods is to construct discriminative spatio-temporal representations. However, confounder bias is often introduced in this process, which adversely affects predictions.

6.2.2 Causal Inference

Causal inference [238, 239] is aimed at pursuing the causal effect to mitigate spurious correlations, which has been seen in epidemiology, political and psychological research for years [225, 240]. Recent years have also witnessed growing applications of causalities in visual tasks such as long-tail classification [225, 226], image semantic segmentation [227], visual dialog [229], image captioning [120] and Visual Question Answering [117, 230, 231]. Beyond applications in image-level vision tasks, causal inference has also attracted increasing attention in video-related tasks. Yang et al. [120] proposed a causality-inspired framework DCM for video moment retrieval to address temporal location biases and improve query-video matching. D2Q [241] is developed to address duration bias in watch-time prediction by leveraging causal reasoning for unbiased video recommendations. Wang et al. [242] presented DeCalGAN with causal inference to generate unseen action features for zero-shot action recognition, addressing confounders effectively. Inspired by visual commonsense learning [243], this is, to the best of our knowledge, the first study to explore confounding effects through causal inference in surgical video understanding, applicable across multiple granular tasks.

6.3 Methodology

6.3.1 Overview

The overall framework of DSVRL is illustrated in Figure 6.2 for achieving a unified spatio-temporal representation for surgical video understanding at different levels. Specifically, to generate the feature of frame x_t, we extract a video clip as the model input, which contains the current frame x_t and a set of its previous frames, as $x = x_{t-T}, ..., x_{t-1}, x_t$, with $T+1$ frames.

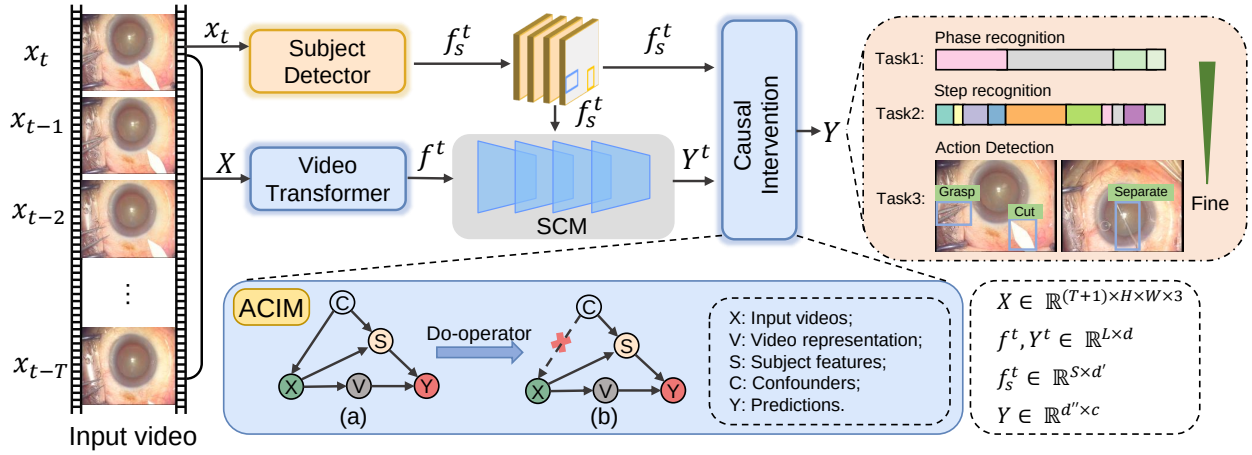


Figure 6.2: The pipeline of our Deconfounded Surgical Video Representation Learning (DSVRL) framework. Given an input video with $T+1$ frames, DSVRL uses a transformer model and subject detector to extract corresponding features, which are then fed into the Subject-Centric Module (SCM) to capture subject-centric video representation. After that, the surgical Activity Causal Intervention Module (ACIM) is utilized to mitigate the effect of confounders, revealing the true causal relationships. The representations learned by the DSVRL framework can be utilized for three different granularities of video understanding tasks.

In this work, we serve the surgical instruments as subjects. The clip is first passed forward a transformer backbone to generate feature maps f_t. After that, an SCM is proposed to refine these video features with subject-specific features f_b^t, focusing on the spatial-temporal relationships between subjects and their surroundings.

Furthermore, a key innovation of the framework is the incorporation of causal inference to address confounding effects that may lead to misleading correlations. The model uses a causal graph to explicitly represent these relationships, where the confounders C impact both inputs X and Y. The confounding effect introduces the harmful spurious correlation between X and Y via the backdoor path: $X \leftarrow C \rightarrow S \rightarrow Y$, which the framework mitigates through a deconfounding mechanism ACIM. This ensures that the model's predictions are

based on causal relationships rather than spurious correlations.

Finally, the refined features Y are passed through a transformer-based decoder that generates multi-granularity predictions, including phase recognition, step recognition, and action detection. The proposed network with SCM and ACIM enables the model to accurately recognize surgical activities (phases, steps or actions) while avoiding biases introduced by confounders.

6.3.2 The Video Transformer Architecture

As illustrated in Figure 6.2, the video transformer as encoder ingests a video clip X of RGB frames $X = (x_{t-T}, ..., x_{t-1}, x_t)$ with $T + 1$ frames, where $x_i \in \mathbb{R}^{H \times W \times 3}$. The input is first "patchified" with temporally extended 3-D patches instead of 2-D image patches to produce a down-sampled tensor $X_p \in \mathbb{R}^{(T+1) \times H' \times W' \times C}$. The tensor appended with a learnable classification token undergoes processing through self-attention layers to derive spatio-temporal visual representations denoted as $f^t \in \mathbb{R}^{L \times d}$. To model the complex temporal cues of the surgical scene, our work uses the MViTv2 [211] model as the encoder. It allows capturing simple low-level visual information at early layers and spatially coarse but complex and high-dimensional features at deeper layers, benefiting from its multi-scale feature hierarchies.

6.3.3 Subject-Centric Video Representation

To get a set of refined tokens $Y^t \in \mathbb{R}^{L \times d}$ contextualized with instrument subject-centric information, we propose a transformer-based Subject-Centric Module (SCM). The SCM takes two inputs: the first is the output of the preceding transformer block, $f^t \in \mathbb{R}^{L \times d}$, which encodes spatio-temporal features of the video. The second input consists of the bounding boxes and corresponding appearance features at time t for subjects, represented by $S^t \in \mathbb{R}^{S \times 4}$

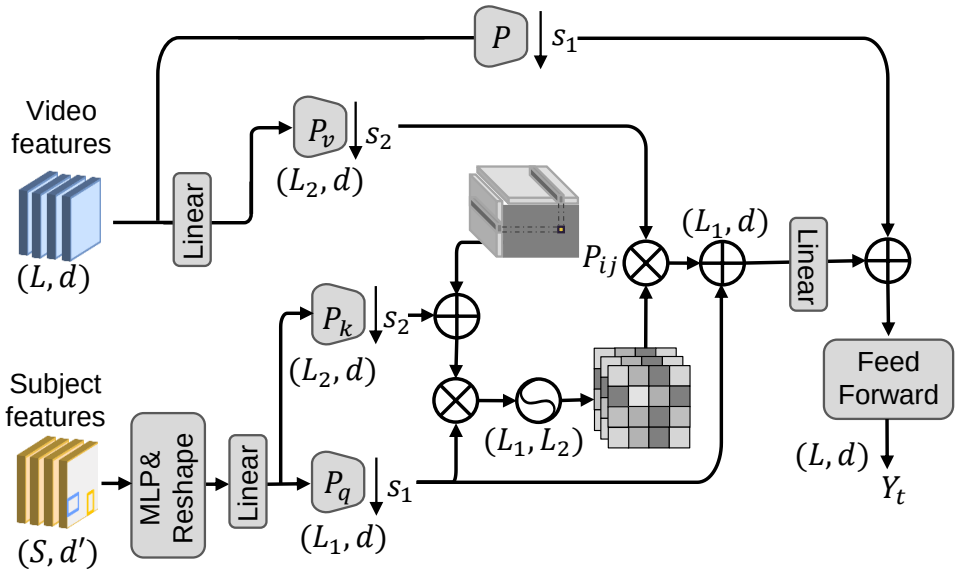


Figure 6.3: Detailed structure of the Subject-Centric Module (SCM).

and $\hat{f}_s^t \in \mathbb{R}^{S\times d'}$, respectively. The key cues for activity recognition in videos are derived from the interaction of subject-related information across both space and time. To capture these cues, we operate the following subject-level stream. For the subject-specific features $\hat{f}_s^t$, we first propose an architecture that includes an MLP and reshape operator, ensuring that the dimensions are consistent with the video features f^t to get $f_s^t \in \mathbb{R}^{L\times d}$, enabling subsequent attention calculations. Then, we apply pooling operations to both f^t and f_s^t, performing spatio-temporal downsampling with scale factors s_1 and s_2, respectively. This process yields the query (f_q^t), key (f_k^t), and value (f_v^t) tensors, which are then used for attention computation.

$$f_q^t = P_q(f_s^t W_q; s_1), f_k^t = P_k(f_s^t W_k; s_2), f_v^t = P_v(f^t W_v; s_2) \tag{6.1}$$

where $f_q^t \in \mathbb{R}^{L_1\times d}$ and $f_k^t, f_v^t \in \mathbb{R}^{L_2\times d}$. $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ represent the projection matrices. As the subjects (instruments) in surgical scenes conduct motions and activities, we intend to select value video features in space and time that are highly active with the subject features and merge them. Spatial-temporal subject-centric attention matrix $Z^t \in \mathbb{R}^{L_1\times d}$ is then obtained as follows. The video representation can be enhanced under the guidance of

query f_q^t. To ensure the shift-invariance in vision, the relative position embedding R^t [211] is added to the computation.

$$\begin{aligned} Z^t &= Softmax((f_q^t(f_k^t)^T + R^t)\sqrt{d})f_v^t, \\ R^t &= \sum\nolimits_{ij} R_{ij}^t = \sum\nolimits_{ij}(f_{qi}^t \cdot P_{ij}^t), \\ R_{ij}^t &= f_{qi}\left(P_{h(i),h(j)}^h + P_{w(i),w(j)}^w + P_{t(i),t(j)}^t\right) \end{aligned} \tag{6.2}$$

where $\sqrt{d}$ is employed to standardize the inner product matrix row-wise. R_{ij}^t is computed based on the spatial-temporal position of elements i and j. Specifically, it is derived from the summation of position embeddings across the height, width, and temporal dimensions, represented as P^h, P^w, and P^t correspondingly. Here, $h(\cdot)$, $w(\cdot)$, and $t(\cdot)$ indicate the vertical, horizontal, and temporal positions of the token, respectively.

To enhance the flow of subject information even more, we utilize the residual video feature pooling connection: $Y^t = Z^t + P(f^t; s_1)$. This helps us obtain the ultimate refined token Y^t, enriched with subject-centric details.

Unlike the current methods that directly use video representation for surgical understanding, we make full use of the activity key cues. With dominant subject features and auxiliary context information, the SCM can effectively model all the activity representations over time. It enables the highlighting of the spatio-temporal regions that interact with the subjects, which includes surrounding tissues and movement information. In this work, we extract the subject-specific features f_t^s from a DINO [213] model. To mitigate the significant domain shift from the natural images on which DINO was originally trained, the model was first pre-trained on a large corpus of public surgical instrument datasets (including CaDIS dataset [155], EndoVis 2017 [152], and EndoVis 2018 [153]), as further detailed in Section 6.4.3. This domain-specific adaptation makes it extensively optimized to ensure that our framework receives accurate subject guidance right from the outset.

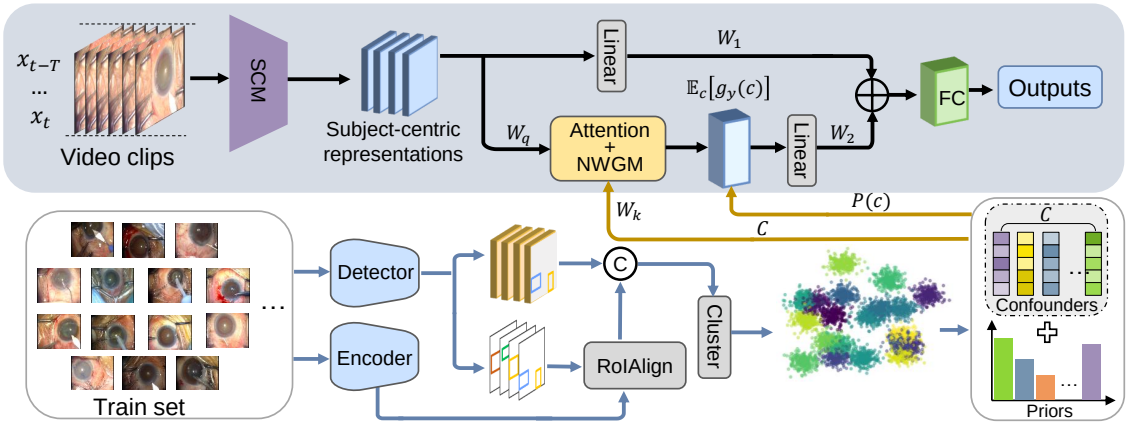


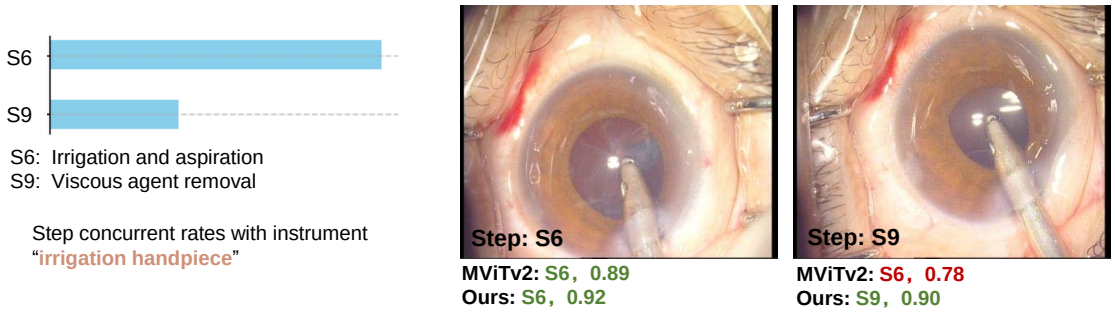
Figure 6.4: Detailed structure of the Activity Causal Intervention Module (ACIM).

6.3.4 Surgical Activity Causal model

6.3.4.1 Causal View at Surgical Scene Understanding

As discussed in the Introduction section, confounds can confuse the surgical activities, result in a sub-optimal solution and degrade the model generalization. To further elaborate on how the bias misleads the results, we build a structured causal model shown in Figure 6.2. It is a directed acyclic graph that can be paired with data to produce quantitative causal estimates. The nodes in the graph denote causal entities, and the edges denote causalities between two nodes: $V \rightarrow Y$ means effect Y is caused by V.

As shown in Figure 6.2, the causal graph involves five variables: input videos X, input video representation V, confounder C, subject features S and prediction Y. Video representation V is obtained from the transformer with X. Traditional surgical understanding models [64, 95, 100, 200] have only one link: $X \rightarrow V \rightarrow Y$, and predict the target by probability $P(Y|X)$. In our causal graph, the subject appearance is a confounder that influences both the correlated variable Y and independent variable X, leading to a spurious correlation between X and Y. The details of the causal relationships are described below.



(a) Step occurrence rates **(b) Qualitative results comparison**

Figure 6.5: Examples about the spurious correlation in step recognition. (a) shows the step concurrent rates with the instrument "irrigation handpiece". (b) shows a comparison of the qualitative results for the MViTv2 [211] and ours on the different frames. The predicted correct and incorrect results are colored green and red, respectively. Due to highly similar visual features and data bias, the MViTv2 [211] learns the spurious correlation in these cases and gives wrong results. Thanks to causal inference, ours achieves more accurate predictions.

$C \rightarrow S \rightarrow Y$. The nature of surgical video data leads to a long-tailed distribution of activity annotations. It introduces a harmful causal effect that misleads the prediction of activities at the tail biased towards the data-rich head. In surgical situations, a single instrument subject frequently serves for multiple actions or steps or phases. During the model's learning process, the subject's appearance will be bound to the corresponding data-rich activity class, forming a certain stereotype.

$C \rightarrow X$. It denotes the effect of C on input videos. In surgical understanding, the subject appearance in videos leads to a strong spatial prior that affects the representation learning of activities in videos. The video spatial-temporal representation extracted from the video encoder is usually entangled with a latent instrument subject appearance factor. Although it enriches the representation with subject-aware features, it introduces the harmful spurious correlation between X and Y via the backdoor path: $X \leftarrow C \rightarrow S \rightarrow Y$, i.e., overexploiting the co-occurrence between the subject appearance and activity class labels to learn biased

visual representations of videos.

$X \rightarrow S \rightarrow Y$ and $X \rightarrow V \rightarrow Y$. The desired effect of the input videos X on the activity prediction Y in our model is captured through two distinct causal paths: $X \rightarrow S \rightarrow Y$ and $X \rightarrow V \rightarrow Y$. In this causal graph, S represents the integration of the subject features, which are derived from subject detectors. In practice, both S and V are typically integrated to jointly contribute to the final prediction. This integration is often realized through methods such as correlation loss or attention mechanisms [105, 236]. In our work, we develop SCM to achieve it, which enables the model to optimally leverage both contextual and subject-specific information, ultimately ensuring the most accurate prediction of the activity Y.

According to the causal theory [109], the confounder C is the common cause of the input video X and corresponding predictions Y. The positive effects of subject-guiding spatial-temporal features providing valuable semantics follow the causal paths $X \rightarrow V/S \rightarrow Y$, which we aim to achieve. Unfortunately, the confounder C causes the negative effect of misleading the model to focus on spurious correlations instead of pure causal relationships. This adverse effect follows the backdoor causal path $X \leftarrow C \rightarrow S \rightarrow Y$, making the effect of truly valuable visual and dynamic content largely overlooked. Next, we propose a deconfounding method to remove the confounding effects.

6.3.4.2 Causal Intervention via Backdoor Adjustment

As shown in Figure 6.2, existing surgical understanding methods rely on the likelihood $P(Y|X)$, which is usually affected by the confounder C and gives rise to spurious correlations. To see this, by applying the Bayes rule on the causal graph, we have:

$$P(Y|X) = \sum\nolimits_{c} P(Y|X, V = f_v(X), S = f_s(X, c))P(c|X) \tag{6.3}$$

where $f_v(\cdot)$ and $f_s(\cdot)$ are two generalized encoding functions that obtain the total V and S, respectively. The confounder C introduces the observational bias via $P(c|X)$. For instance in Figure 6.5 (a)(b), in the unbalanced surgical dataset, $P(c = \textit{irrigation handpiece}|X = S6)$ is large while $P(c = \textit{irrigation handpiece}|X = S9)$ is small. According to Eq. (6.3), $P(Y = S6|X = S9, V = S9, S = \textit{irrigation handpiece})$ plays a more important role than $P(Y = S9|X = S9, V = S9, S = \textit{irrigation handpiece})$ in estimating $P(Y|X)$. $S9$ and $S6$ denote the class label of the surgical step. Thus, the classifier learns the correlation between the visual features of "irrigation handpiece" and the class label of $S6$ by mistake, i.e., the learned video features of the $S6$ are actually the instrument "irrigation handpiece" like visual features, without truly looking into the details of more important video temporal context information.

Although the confounder misleads the model prediction severely, it on the other hand is indispensable for the training. Thanks to the causal theory [109], we can realize the deconfounded training through backdoor adjustment without changing these useful components:

$$P(Y|do(X)) = \sum\nolimits_c P(Y|X, V = f_v(X), S = f_s(X, c))P(c) \tag{6.4}$$

Causal intervention $P(Y|do(X))$ is introduced into our task to block the backdoor path $X \leftarrow C \rightarrow S \rightarrow Y$, where the do calculus $do(\cdot)$ plays a role of cutting off the edge $C \rightarrow S \rightarrow Y$, e.g., cut off the correction that S_6 relies on "irrigation handpiece". Compared to Eq. (6.3), c is no longer affected by X, and thus the intervention deliberately forces X to incorporate every c fairly, subject to its prior $P(c)$, into the prediction of Y. In this way, the proposed model removes the confounding effect and learns the true causality from X to Y, leading to high-quality video representations.

6.3.4.3 Interventional Surgical Scene Understanding

To implement the theoretical and imaginative intervention in Eq. (6.4), we propose a causal intervention architecture to achieve the deconfounded training for our model. Given the video representation and subject features, Eq. (6.4) is implemented as $\sum_c P(y|x, v = f_v(x), s = f_s(x, c))P(c)$, where $P(y|x, v = f_v(x), s = f_s(x, c))$ is output of our proposed method $f_y(\cdot)$.

$$P(Y|do(X)) := \mathbb{E}_c[\sigma(f_y(v, i, c))] \tag{6.5}$$

where $\sigma(\cdot)$ is the softmax function that transforms the output $f_y(\cdot)$ into $[0, 1]$. When calculating Eq. (6.5), we find it requires multiple forward passes of all c, and the computational overhead is expensive. Fortunately, by applying the Normalized Weighted Geometric Mean (NWGM) approximation [243, 244], we approximately move outer expectation $\mathbb{E}_c$ into the softmax function:

$$\mathbb{E}_c[\sigma(f_y(v, i, c))] \overset{NWGM}{\approx} \sigma(\mathbb{E}_c[f_y(v, i, c)]) \tag{6.6}$$

Inspired by [243], the effect on Y comes from both V, S and confounder C (as shown in Figure 6.2). We parameterize a network model to approximate the above conditional probability of Eq. (6.6) as follows:

$$\mathbb{E}_c(f_y(v, i, c)) = W_1 h + W_2 \cdot \mathbb{E}_c(g_y(c)) \tag{6.7}$$

where the parameters $W_1 \in \mathbb{R}^{d_m \times d}$ and $W_2 \in \mathbb{R}^{d_m \times d_c}$ are trainable and $h = \alpha(v, s) \in \mathbb{R}^{d \times 1}$. Here, $\alpha(\cdot)$ represents our SCM that combines v and s into the joint representation h. To calculate $\mathbb{E}_c(g_y(c))$, scaled dot-product attention [245] is utilized to assign weights for different confounders in dictionary C with specific h. Specifically, given the h and confounder dictionary C, we can have:

$$\mathbb{E}_c(g_y(c)) = \sum\nolimits_c [softmax(\frac{(W_q h)^T (W_k c)}{\sqrt{d_c}}) \odot c]P(c) \tag{6.8}$$

The embedding matrices are denoted as $W_q \in \mathbb{R}^{d_n \times d}$ and $W_k \in \mathbb{R}^{d_n \times d_c}$, which is employed to map each vector to the common subspace for similarity measure. $P(c)$ denotes the prior statistic probability and $\odot$ is the element-wise product. The result of scaled dot-product attention is a weight coefficient that measures the importance of each c after interacting with the origin feature h. d_c denotes a constant scaling factor.

For the confounder set C, since we can hardly enumerate all the subject appearances in the surgical scenarios, we approximate it as a stratified confounder dictionary $C = [c_1, c_2, ..., c_N]$, where N is a hyperparameter indicating the size, and each $c_i \in \mathbb{R}_c^d$ represents a prototype of subject appearance. As shown in Figure 6.4, in diverse surgical scenes, we train specific subject detectors based on DINO [213]. By utilizing pre-trained subject detectors, we acquire bounding box sets of subjects and corresponding features for various surgical scenes. With the same bounding boxes, RoIAlign is applied to extract the subject region features based on the video feature map. To further highlight the effect of subject appearance in our work, the Regions of Interest (RoI) features are integrated with subject-specific features from detectors to form $H = \{h_k \in \mathbb{R}_c^d\}_{k=1}^{N_h}$, where N_h is the number of training samples. To obtain c_i, we applied clustering algorithm to H to obtain different clusters c_i. K-Means++ with principle component analysis is utilized in our work for efficiency and effectiveness. Due to the long-tailed data distribution, each cluster has varying features. Each cluster c_i is determined as the average feature of each cluster:

$$c_i = \frac{1}{N_i} \sum_{j=1}^{N_i} h_{ij} \tag{6.9}$$

where N_i is the number of subject appearance features in the i-th cluster. We can also get $P(c_i) = \frac{N_i}{N_h}$.

6.3.5 Overall Training Objective

As a training objective, we employ a cross-entropy loss for surgical understanding. Given the activity prediction $Y = \{y_t\}_{t=1}^{T}$ from our framework and the target labels $\hat{Y} = \{\hat{y}_t\}_{t=1}^{T}$, the objective can be formulated as:

$$L_{ce} = -\frac{1}{T}\sum_{t} \hat{y}_t \log(y_t) \tag{6.10}$$

6.4 Experiments

6.4.1 Datasets

The proposed pipeline is evaluated on three tasks using four surgical video datasets, including the public datasets GraSP, CATARACT, and AutoLaparo, as well as our private cataract dataset, Cataract_ours. Brief descriptions of these datasets are provided below, and the overview of the four datasets is shown in Table 6.1.

Cataract_ours dataset consists of 61 videos (totalling 85659 frames) with a frame rate of 1 frame per second (fps) and a resolution of 720 × 576. For phase and step recognition tasks, each frame is annotated with corresponding phase and step labels by ophthalmologists, covering 5 phases and 11 steps. In our experiments, we randomly split the dataset into a training set with 35 videos (51510 frames), a validation set with 6 videos (4874 frames), and a test set with 20 videos (29275 frames). For action detection, as in [163], we manually annotate 20 videos (23580 frames and 28428 action instances), labelling each frame with 49 action labels, 13 instrument labels and their corresponding bounding boxes. The dataset is then randomly divided into a training set of 12 videos (15602 frames with 18875 instances), a validation set of 3 videos (2039 frames with 2263 instances), and a test set of 5 videos

Table 6.1: Overview of four datasets

Dataset	Videos	Frames	Resolution	Task
Cataract_ours	61	85659	720×576	Phase, Step, Action
GraSP	13	116521	1280×800	Phase, Step, Action
CATARACTS	50	17077	1920×1080	Phase, Step
AutoLaparo	21	40211	1920×1080	Phase

(5939 frames with 7290 instances) for this task.

GraSP dataset [20] includes 13 robotic-assisted laparoscopic radical Prostatectomy videos, performed with the Da Vinci SI 3000 Surgical System manufactured by Intuitive Inc. (Sunnyvale, California). It includes 11 phases, 21 steps, 14 actions and 7 instruments, and the resolution is 1280 × 800. The phase and step annotations are provided by frames at one fps, while the actions and instruments are given by frames every 35 seconds of video. Thus, 116521 frames are annotated with one phase-step pair, while 3449 have instrument and action annotations. We split the dataset following work [20], including a training set with 8 videos (73618 frames) and a test set with 5 videos (42903 frames) for phase and step recognition, as well as a training set with 8 videos (2324 frames) and a test set with 5 videos (1125 frames) for action detection.

CATARACTS dataset [41] consists of 50 videos of cataract surgeries performed by three surgeons at Brest University Hospital. The frame resolution is 1920 × 1080 pixels, with a frame rate of 1 fps. The dataset includes 19 steps. Based on the step labels provided by the dataset and the annotation requirements of Cataract_ours, we further expand the phase labels of the CATARACTS dataset, including 5 phases, consistent with Cataract_ours. Following the split in [41], the dataset is divided into a training set with 25 videos (17077 frames), a validation set with 5 videos (3728 frames) and a test set with 20 videos (12250

frames).

AutoLaparo dataset [87] contains 21 videos of laparoscopic hysterectomy collected from Prince of Wales Hospital, Hong Kong. The videos are downsampled to 1 fps with a standard resolution of 1920 × 1080 pixels. The hysterectomy procedure is divided into 7 phases in the dataset, and each frame is annotated with a phase label. Following the split in [87], 21 videos are divided into three subsets for training, validation, and test, containing 10 videos (40211 frames), 4 videos (14972 frames), and 7 videos (28060 frames), respectively.

6.4.2 Evaluation Metrics

For the tasks of phase and step recognition, the performance is evaluated using four commonly utilized metrics: accuracy, precision, recall, and Jaccard index. Accuracy represents the percentage of correctly classified frames across the entire video, while precision, recall, and Jaccard are calculated as follows:

$$\text{PR} = \frac{|GT \cap P|}{|P|}, \quad \text{RE} = \frac{|GT \cap P|}{|GT|}, \quad \text{JA} = \frac{|GT \cap P|}{|GT \cup P|} \tag{6.11}$$

where GT represents the ground truth set and P denotes the prediction set, respectively, while PR, RE and JA denote the precision, recall, and Jaccard. The accuracy reflects the proportion of correctly identified frames within a video and is video-based. Nevertheless, due to class imbalance in the video, the influence of shorter phases or steps on overall accuracy is minimal. To provide a more comprehensive evaluation of our model, we further employ class (phase or step)-specific precision, recall, and Jaccard indices, which correspond to positive predictive value, sensitivity, and the intersection-over-union between the predicted outcomes and GT, respectively.

To evaluate the performance of action detection, we employ mean Average Precision (mAP), a widely used metric for action detection in videos. We compute mAP at Intersection-

over-Union (IoU) thresholds of 0.5 ($mAP_{0.5}$), 0.75 ($mAP_{0.75}$), and 0.95 ($mAP_{0.95}$), and then calculate their average (Avg.).

6.4.3 Implementation Details

The model is developed using Pytorch [182] and trained on an NVIDIA RTX A6000 GPU. We utilize the AdamW optimizer with an initial learning rate of $1e-3$ and a weight decay of 0.05. A cosine learning rate scheduler is employed, with a linear warm-up period of 10 epochs, starting at a learning rate of $1e-6$. Data augmentation techniques, including 224×224 cropping, jittering, and flipping, are applied to enhance the training dataset. Following the approach of [211], the sequence length is set to 16 frames.

To obtain subjects-related information in the surgical videos, we first train a surgical instrument detector. Based on the CaDIS dataset [155], along with the EndoVis 2017 [152] and EndoVis 2018 [153] datasets, the instrument detector is pre-trained using the DINO-5scale [213] with a Swin-L [246] backbone. The trained DINO-5scale model is utilized to predict the box-specific features and bounding boxes required for each dataset. We select bounding boxes with a confidence threshold of 0.75.

6.4.4 Comparison with state-of-the-arts

Tables 6.2-6.9 present the results of our proposed framework compared to the current SOTA methods across three tasks on four surgical datasets. The detailed analysis is as follows.

6.4.4.1 Action Detection

The action detection task is the most challenging among the three tasks due to the wide variety of action classes and the possibility of multiple different actions occurring in a single frame. This makes learning temporal information especially demanding. In the action detection task, we compare the most advanced action detection methods on the Cataract_ours and GraSP datasets: ACTNet [163], X3D [247], Rev-MViT [248], TAPIR [21], and MViTv2 [211].

ACTNet [163] is an anchor-context action detection network that predicts action locations, classes, and confidence scores. X3D [247] is a highly efficient spatio-temporal architecture derived from progressively expanding a compact spatial network. Rev-MViT [248] represents memory-efficient reversible adaptations of Vision Transformers (ViT) and Multiscale Vision Transformers (MViT). TAPIR [21] serves as a baseline framework designed for different levels of task. MViTv2 [211], the baseline of our framework, is a transformer-based architecture tailored for video recognition by utilizing multiscale feature hierarchies.

For the Cataract_ours dataset, we train DSVRL on the training set, validate it on the validation set to select the best model, and finally report the results on the testing set. As seen in Table 6.2 , our method outperforms all other SOTA methods on the Cataract_ours dataset, surpassing the best transformer-based method MViTv2 by 2.89% in terms of average mAP.

For the GraSP dataset, following the split in [20], we train on the training set and test the model on the last epoch. As shown in the table, our approach significantly outperforms other methods, including both CNN-based and transformer-based models. This strongly demonstrates the effectiveness of the proposed module in the action detection task. Notably, compared to the results on phase and step tasks, the greatest improvement in our framework

is observed in the most challenging action detection task. This aligns with our understanding that subject information and action-related dynamics are key factors in action detection, which are well captured by our proposed module with causal inference.

Table 6.2: Performance comparison with other state-of-the-art methods on various datasets for action detection. The best results are highlighted in bold.

Dataset	Method	$\mathbf{mAP}_{0.5}$	$\mathbf{mAP}_{0.75}$	$\mathbf{mAP}_{0.95}$	Avg.
Cataract_ours	ACTNet [163]	39.23	37.22	15.31	30.59
	X3D [247]	44.71	42.47	19.07	35.42
	Rev-MViT [248]	41.00	38.87	17.12	32.33
	TAPIR [21]	52.28	49.92	23.65	41.95
	MViTv2 [211]	55.69	53.33	24.80	44.61
	Ours	**59.16**	**56.77**	**26.57**	**47.50**
GraSP	ACTNet	24.22	22.41	6.00	17.53
	X3D	25.42	23.04	5.45	17.97
	Rev-MViT	24.57	22.43	6.31	17.77
	TAPIR	26.59	24.57	5.90	19.02
	MViTv2	28.11	26.13	7.20	20.48
	Ours	**30.13**	**27.47**	**8.24**	**21.95**

6.4.4.2 Step Recognition

The goal of step recognition is to identify different steps of a surgical video that are directly related to the instruments, which are the subjects involved in the surgery. In this task, we evaluate the performance of our model based on three datasets: Cataract_ours, GraSP [20], and CATARACT [41]. We also compare our model with the SOTA methods for step

Table 6.3: Performance comparison with other state-of-the-art methods on Cataract_ours dataset for step recognition.

Method	**Frame**	**Video-level**	**Step-level**		
	Length	**Accuracy**	**Precision**	**Recall**	**Jaccard**
NETE [201]	Video	90.1±11.1	86.8±9.7	85.8±6.4	76.0±10.4
SAHC [95]	Video	88.8±10.5	81.8±11.5	85.1±6.2	71.8±11.5
TeCNO [64]	30	87.5±10.5	82.1±10.8	80.1±10.9	68.5±13.1
TMRNet [200]	30	88.1±9.0	83.0±11.7	78.7±11.1	68.5±13.7
Trans-SVNet [235]	30	85.2±10.3	80.1±11.7	72.7±17.4	62.4±16.9
MViTv2 [211]	16	91.1±6.8	85.2±12.0	84.2±11.0	73.4±13.7
Ours	16	**93.0±4.3**	**88.9±6.4**	**87.9±6.7**	**79.5±9.2**

Table 6.4: Performance comparison with other state-of-the-art methods on CATARACTS dataset for step recognition.

Method	**Frame**	**Video-level**	**Step-level**		
	Length	**Accuracy**	**Precision**	**Recall**	**Jaccard**
NETE	Video	77.1±13.8	56.9±31.0	53.6±28.1	42.4±25.3
SAHC	Video	71.5±12.2	45.4±31.3	43.7±31.4	32.8±26.1
TeCNO	30	80.4±11.9	56.6±34.3	51.6±35.0	40.4±28.1
TMRNet	30	81.2±10.6	70.1±32.4	50.7±35.5	42.3±29.6
Trans-SVNet	30	76.3±10.8	51.3±32.1	47.2±33.1	37.4±27.5
MViTv2	16	87.6±7.9	68.6±32.0	68.2±31.5	58.0±28.2
Ours	16	**88.0±7.7**	**71.6±28.4**	**71.0±30.3**	**60.5±27.2**

Table 6.5: Performance comparison with other state-of-the-art methods on GraSP dataset for step recognition.

Method	**Frame**	**Video-level**	**Step-level**		
	Length	**Accuracy**	**Precision**	**Recall**	**Jaccard**
NETE	Video	50.6±8.98	38.0±30.2	32.6±28.0	20.4±18.8
SAHC	Video	51.8±8.85	30.3±26.5	28.2±29.2	17.4±19.0
TeCNO	30	52.3±6.0	28.7±26.0	32.6±31.7	18.6±19.1
TMRNet	30	55.3±7.6	44.9±21.0	41.2±25.0	26.5±16.1
Trans-SVNet	30	55.6±6.4	39.5±22.3	36.6±28.2	23.4±18.5
MViTv2	16	62.3±9.9	49.6±16.4	54.1±29.1	33.0±18.0
Ours	16	**63.4±10.0**	**51.9±14.6**	**53.2±24.0**	**34.0±15.5**

recognition: NETE [201], SAHC [95], TeCNO [64], TMRNet [200], Trans-SVNet [235], and MViTv2 [211].

NETE [201] divides the end-to-end training into an initial prediction phase and a refinement phase, enabling more efficient convergence and optimization. Building on NETE, SAHC [95] is recently introduced to learn high-level segments from surgical videos, helping to resolve ambiguities caused by low-level frame errors. The TeCNO network performs workflow recognition using a MS-TCN [234]. TMRNet [200] employs a memory bank to capture long-sequence temporal information, utilizing even 30-second frame sequences in our comparison. Trans-SVNet [235] introduces a novel hybrid embedding aggregation transformer for workflow recognition in surgical videos.

Table 6.3, 6.4 and 6.5 compare our method with SOTA approaches in both video-level and step-level metrics. The standard deviation of these scores for video-level metrics reflects variations in performance across different videos, while the standard deviation for step-level

metrics reflects performance variability across different step classes. From Table 6.3, 6.4 and 6.5, it is clear that our method significantly outperforms other SOTA methods in both video-level metrics (accuracy) and step-level metrics (precision, recall, and Jaccard) in the three datasets.

Specifically, for Cataract_ours, MViTv2 is the best-performing existing method in terms of accuracy, but our model exhibits a higher performance, with a 1.9% improvement in accuracy. DSVRL also surpassed NETE, the champion model for step recognition, by 2.1% in precision, 2.1% in recall, and 3.5% in Jaccard. Additionally, the tables show that while other methods require up to 30 frames of historical data or even the entire video memory to achieve their results, our method only requires a sequence length of 16 frames, laying the groundwork for more efficient real-world applications.

Notably, the standard deviation of our method is generally lower than that of other approaches across all three datasets, whether in video-level or step-level metrics. This indicates that our method, which utilizes causal relationships for step recognition, can more effectively enhance the performance of low-sample classes and significantly mitigate the adverse effects of class imbalance.

6.4.4.3 Phase Recognition

For the phase recognition task, we validate our model on four datasets: Cataract_ours, GraSP [20], CATARACT [41], and AutoLaparo [87], as shown in Table 6.6, 6.7, 6.8 and 6.9. Although the number of phase classes is smaller than that of step classes, they are still directly associated with the instruments, making our method equally effective in phase recognition. For the Cataract_ours, GraSP [20], and CATARACT [41] datasets, we use the same comparative methods as in step recognition. Our method consistently outperforms the existing SOTA methods, particularly in phase-level metrics.

Table 6.6: Performance comparison with other state-of-the-art methods on Cataract_ours dataset for phase recognition.

Method	Frame	Video-level	Phase-level		
	Length	Accuracy	Precision	Recall	Jaccard
NETE	Video	87.6±10.5	80.9±11.9	78.5±19.3	68.0±20.6
SAHC	Video	86.0±12.7	78.5±14.9	79.4±14.4	67.5±19.6
TeCNO	30	87.6±10.9	81.1±12.5	82.5±10.9	70.8±16.3
TMRNet	30	87.8±7.1	79.7±14.7	79.8±14.3	68.6±19.4
Trans-SVNet	30	86.7±11.4	80.1±13.6	82.2±12.0	69.7±16.8
MViTv2	16	90.4±6.5	83.5±11.5	82.4±15.2	72.8±19.0
Ours	16	**91.8±5.8**	**85.2±11.1**	**85.0±13.4**	**75.6±17.6**

Table 6.7: Performance comparison with other state-of-the-art methods on CATARACTS dataset for phase recognition.

Method	Frame	Video-level	Phase-level		
	Length	Accuracy	Precision	Recall	Jaccard
NETE	Video	81.0±9.3	78.4±12.4	72.8±12.6	60.6±13.5
SAHC	Video	80.8±9.2	74.0±12.0	77.2±9.6	61.3±12.9
TeCNO	30	86.1±6.5	81.4±9.9	83.2±5.4	70.0±9.8
TMRNet	30	85.0±8.0	81.4±7.3	80.2±5.6	67.9±7.7
Trans-SVNet	30	84.3±6.4	81.0±9.4	79.5±6.5	67.0±9.0
MViTv2	16	89.3±9.7	82.3±13.9	85.0±4.3	71.5±11.1
Ours	16	**90.7±8.3**	**85.2±12.8**	**86.7±7.4**	**74.9±10.0**

Table 6.8: Performance comparison with other state-of-the-art methods on GraSP dataset for phase recognition.

Method	Frame	Video-level	Phase-level		
	Length	Accuracy	Precision	Recall	Jaccard
NETE	Video	72.9±8.2	76.4±14.8	68.6±24.0	55.0±22.3
SAHC	Video	73.2±7.5	68.5±18.1	70.1±24.4	54.2±21.4
TeCNO	30	73.0±5.3	69.4±12.7	71.0±20.0	53.6±16.0
TMRNet	30	72.5±5.7	73.7±12.1	68.5±23.1	53.1±16.4
Trans-SVNet	30	72.4±4.0	72.9±9.8	70.3±24.4	53.4±16.9
MViTv2	16	76.1±4.8	75.6± 13.1	71.6±22.6	55.5±17.0
Ours	16	**76.7±4.2**	**76.1±11.1**	**72.5±22.0**	**57.2±16.6**

The comparative results on the AutoLaparo dataset are drawn from [250], where it can be observed that our method significantly surpasses LoViT, with a 1.3% improvement in accuracy, a 7.8% increase in recall, and a 4% improvement in Jaccard score. Additionally, while other methods require a frame length of 30 frames or even the entire video as input, our method achieves superior performance with only 16 frames of historical information. The standard deviation of our method is also consistently lower than that of the comparative approaches, further highlighting its advantage in improving the performance of low-sample classes.

6.4.4.4 Comparison Performance on Few-sample Classes

As shown in Figure 6.6, we visually compare the performance of our method with existing approaches on classes with fewer samples in the phase and step recognition based on the Cataract_ours and GraSP datasets. For example, for phase recognition on the

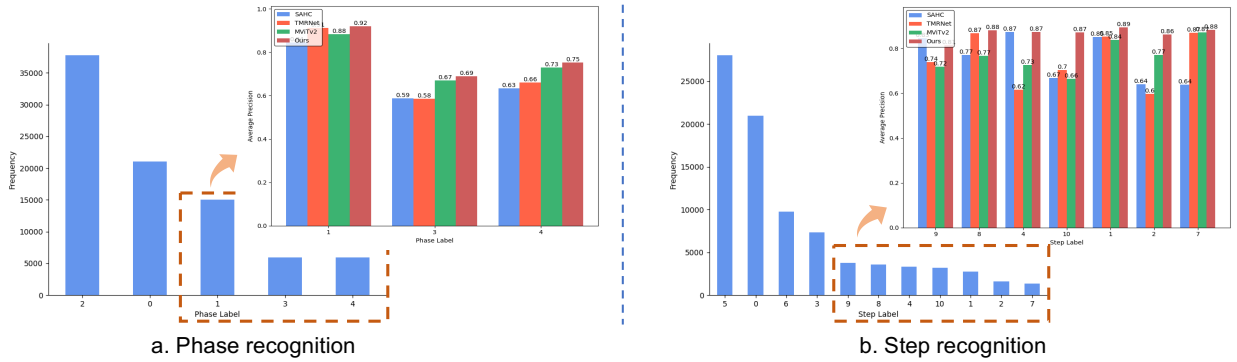
Table 6.9: Performance comparison with other state-of-the-art methods on AutoLaparo dataset for phase recognition.

Method	**Frame**	**Video-level**	**Phase-level**		
	Length	**Accuracy**	**Precision**	**Recall**	**Jaccard**
SV-RCNet [91]	10	75.6	64	59.7	47.2
TMRNet	30	78.2	66	61.5	49.6
TeCNO	30	77.3	66.9	64.6	50.7
Trans-SVNet	30	78.3	64.2	62.1	50.7
AVT [249]	30	77.8±9.4	68	62.2	50.7
LoViT [250]	3000	81.4±7.6	**85.1**	65.9	55.9
MViTv2	16	81.9±9.7	75.4±9.6	72.8±13.1	59.3±11.0
Ours	16	**82.7±9.7**	76.4±11.2	**73.7±14.2**	**59.9±13.3**

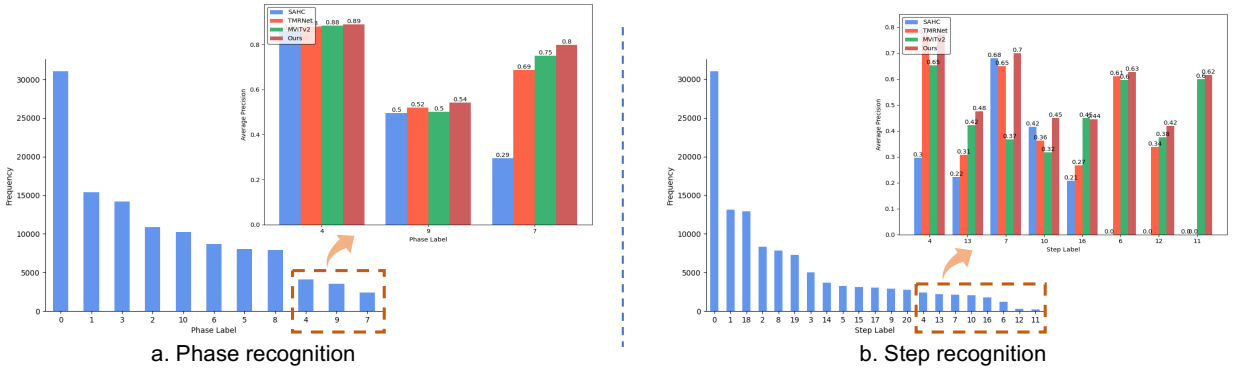
Cataract_ours dataset, we compare our method's performance with that of SAHC [95], TMRNet [200], and MViTv2 [211] for classes 1, 3 and 4, using precision as the evaluation metric (class-level). The results clearly show that our method achieves the best performance.

In step recognition, we select a more significant number of few-shot classes for comparison (7 classes in the Cataract_ours dataset). Although our method does not achieve the best performance in every class, it performs excellently in the vast majority of classes. This indicates that our method demonstrates stable performance improvement in few-shot classes.

Moreover, for step recognition on the GraSP dataset, for the three classes with the fewest samples, classes 6, 12, and 11, both SAHC [95] and TMRNet [200] fail to detect or partially detect these classes. In contrast, our method not only detects these classes but also maintains strong performance in recognition.



(a) Cataract_ours dataset



(b) GraSP dataset

Figure 6.6: Visualization comparison on few-sample classes to quantitatively demonstrate enhanced effectiveness.

6.4.5 Ablation Study

6.4.5.1 Study on Key Modules

We conduct ablation experiments on the three tasks in Cataract_ours to validate the contributions of key modules, including the SCM and ACIM, with the results presented in Table 6.10. The models excluding the SCM and ACIM serve as our baseline models, which are built upon the MViTv2 framework and supplemented with different task-specific losses and output heads. To reduce computational costs, all of our improvements are applied to the final block. We compare the effects of adding the SCM to the baseline model, as well as the combination of the SCM and ACIM.

For action detection, compared to the baseline model, an increase of 1.95% (55.69% v.s. 57.64%), 1.78% (53.33% v.s. 55.11%), and 0.73% (24.80% v.s. 25.53%) in $\text{mAP}_{0.5}$, $\text{mAP}_{0.75}$, and $\text{mAP}_{0.95}$ metrics, respectively, are achieved by introducing the SCM. This highlights the critical role of subject-centric spatio-temporal information in video representation. After further integrating the ACIM, the performance across various tasks is enhanced by significant margins: 1.52% (57.64% v.s. 59.16%) in $\text{mAP}_{0.5}$, 1.66% (55.11% v.s. 56.77%) in $\text{mAP}_{0.75}$, and 1.04% (25.53% v.s. 26.57%) in $\text{mAP}_{0.95}$. This demonstrates that reducing the influence of confounding factors in causal reasoning can significantly boost model performance.

Ablation experiments on phase and step recognition exhibit similar patterns as those observed in the action detection task. In step recognition, after incorporating both key modules into the baseline, our model significantly outperforms the MViTv2 model by 1.52% (57.64% v.s. 59.16%), 1.66% (55.11% v.s. 56.77%), and 1.04% (25.53% v.s. 26.57%) in terms of accuracy, precision, recall, and Jaccard, respectively. For phase recognition, compared to the baseline model, an increase of 1.52% (57.64% v.s. 59.16%), 1.66% (55.11% v.s. 56.77%), and 1.04% (25.53% v.s. 26.57%) is achieved after adding the two key modules in terms of accuracy, precision, recall, and Jaccard, respectively. These also fully demonstrate the effectiveness of SCM and ACIM across the three tasks.

Notably, compared to video-level metrics, the standard deviation in phase or step-level metrics is significantly reduced, especially after adding the ACIM. In the step task, the addition of the ACIM reduce the standard deviation of precision by 5.6%, recall by 4.3%, and Jaccard by 4.5%. In the phase task, while adding the SCM to the baseline improves performance across all four metrics, the standard deviation of precision, recall, and Jaccard also increases after introducing the SCM. However, after incorporating the ACIM, not only is the performance in the phase task improved, but the standard deviation in phase-level metrics is also significantly reduced, even falling below the baseline model's standard deviation by approximately 1.0%. This clearly demonstrates the importance of our model, particularly

Table 6.10: Ablation study on key components for action detection, step recognition, and phase recognition on the Cataract_ours dataset.

Modules		**Action Detection**			
SCM	**ACIM**	**mAP$_{0.5}$**	**mAP$_{0.75}$**	**mAP$_{0.95}$**	**Avg.**
		55.69	53.33	24.8	44.61
✓		57.64	55.11	25.53	45.76
✓	✓	**59.16**	**56.77**	**26.57**	**47.50**
Modules		**Step Recognition**			
SCM	**ACIM**	**Accuracy**	**Precision**	**Recall**	**Jaccard**
		91.1±6.8	85.2±12.0	84.2±11.0	73.4±13.7
✓		91.1±5.2	86.4±10.7	84.7±7.3	74.7±10.7
✓	✓	**93.0±4.3**	**88.9±6.4**	**87.9±6.7**	**79.5±9.2**
Modules		**Phase Recognition**			
SCM	**ACIM**	**Accuracy**	**Precision**	**Recall**	**Jaccard**
		90.4±6.5	83.5±11.5	82.4±15.2	72.8±19.0
✓		91.2±6.3	84.6±12.5	83.4±17.9	73.4±20.2
✓	✓	**91.8±5.8**	**85.2±11.1**	**85.0±13.4**	**75.6±17.6**

the ACIM, in addressing the class imbalance problem. By eliminating the influence of non-causal confounding factors, the performance of minority classes is less affected by the bias toward majority classes in the model.

The t-SNE visualization in Figure 6.7 highlights the impact of our key model components on improving class distinguishability. The blue circle in the figure demonstrates that by adding these components, the model reduces the overlap between classes and enhances the

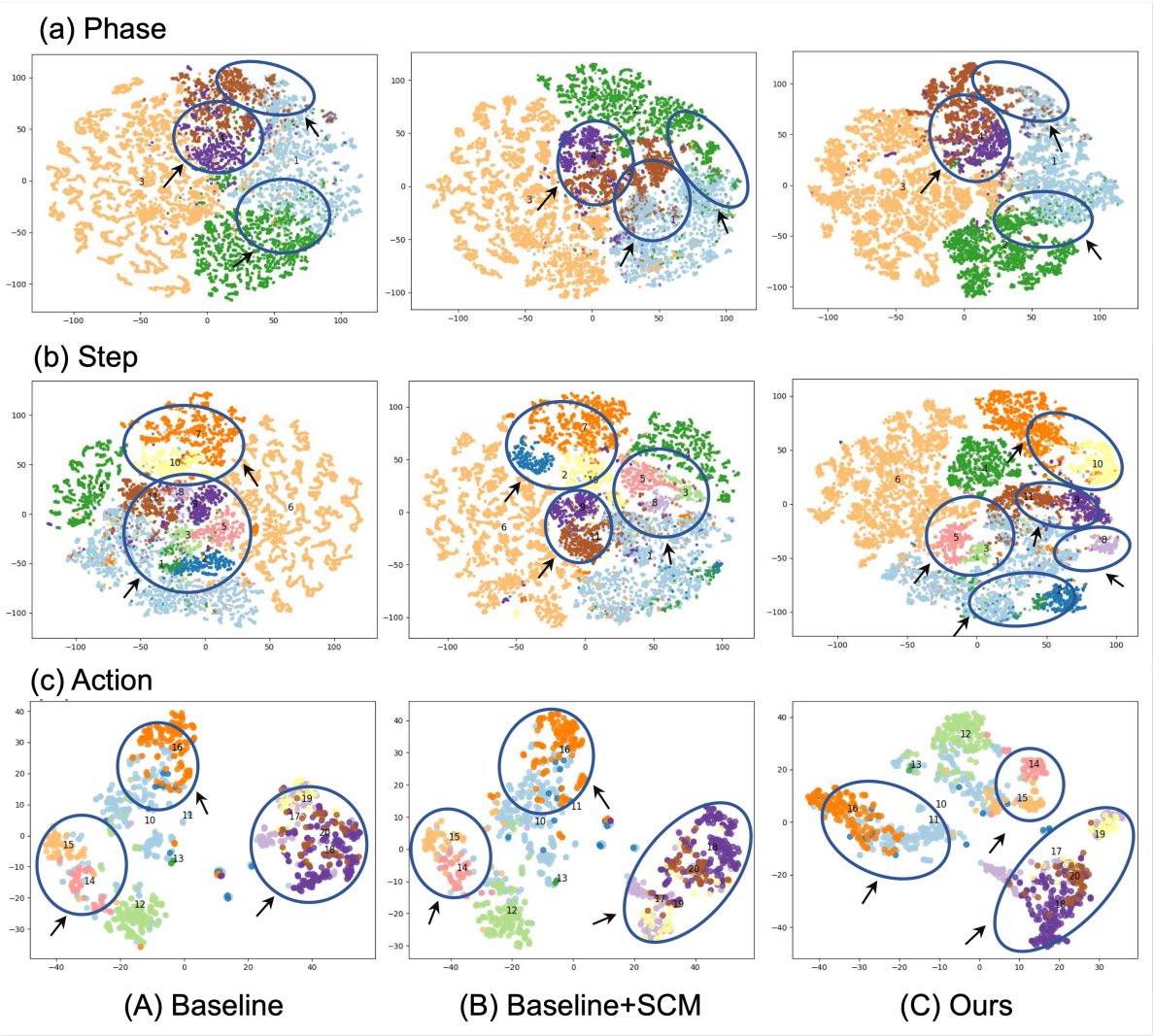


Figure 6.7: T-SNE visualization results of (A) baseline, (B) baseline+SCM and (C) Ours in phase, step and action tasks. The blue circle indicates improvement in effectiveness. Due to space constraints, for action detection, only the distribution of the 11 challenging action classes related to the instrument "cannula" and "capsulorhexis forces" is shown in the figures, illustrating our method's enhanced performance on these difficult action classes.

separation across three different tasks. For action detection, we focus on eleven action classes, which are generated by two different instruments: the cannula, with action classes ranging from 10 to 16, and the capsulorhexis forceps, with action classes spanning from 17 to

20. These action classes naturally exhibit significant spatial similarity and pose classification challenges. In the baseline model (Figure 6.7), these classes are mixed together, making them hard to distinguish. However, after incorporating our proposed modules, the distinction between these classes, as indicated by the blue circle, becomes much clearer.

6.4.5.2 Study on Attention Setting

Table 6.11 shows an ablation study on the impact of attention mechanisms and query types within our SCM, evaluated using $mAP_{0.5}$, $mAP_{0.75}$, and $mAP_{0.95}$, and the average mAP. Using self-attention with only visual features (V, self) yields the lowest performance, with an average mAP of 44.61%, particularly underperforming at higher thresholds like $mAP_{0.95}$. Adding subject and visual features (S+V) with self-attention slightly improves the average mAP to 45.06%, demonstrating the benefit of combining these feature types. Cross-attention mechanisms further boost performance, with visual features as the query (S+V, cross) achieving an average mAP of 45.58%, highlighting its ability to better capture interactions between features. The best results are obtained when subject features are used as the query in the cross-attention setup (S+V, cross), achieving an average mAP of 45.76%, with particularly strong performance at $mAP_{0.75}$. These findings emphasize the effectiveness of cross-attention in integrating subject and visual features to improve performance.

6.4.5.3 Study on Confounder Types

Table 6.12 presents an ablation study evaluating the impact of different confounder types within our ACIM on model performance. The results show that random confounders, which likely introduce noise or less relevant information, yield the lowest performance, with an average mAP of 46.40% and a relatively poor result at $mAP_{0.95}$ (25.55%). Incorporating either ROI features (ROI) or subject features (S) individually improves performance, achieving av-

Table 6.11: Ablation study on attention structure.

Attention	**Query**	**mAP**$_{0.5}$	**mAP**$_{0.75}$	**mAP**$_{0.95}$	**Avg.**
V, self	V	55.69	53.33	24.8	44.61
S+V, self	S	57.01	54.51	23.67	45.06
S+V, cross	V	56.86	54.46	25.44	45.58
S+V, cross	S	**57.64**	**55.11**	**24.53**	**45.76**

V:visual features; S: subject features.

Table 6.12: Ablation study on confounder types.

Confounder	**mAP**$_{0.5}$	**mAP**$_{0.75}$	**mAP**$_{0.95}$	**Avg.**
Random	58.09	55.54	25.55	46.40
ROI	58.27	55.72	**26.72**	46.90
S	58.31	55.77	26.01	46.70
ROI+S	**59.16**	**56.77**	26.57	**47.50**

ROI: roi features; S: subject features.

erage mAPs of 46.90% and 46.70%, respectively, with ROI slightly outperforming subject features at higher thresholds like mAP$_{0.95}$. The best performance is observed when both ROI and subject features (ROI+S) are combined, as shown in Figure 6.4, achieving the highest average mAP of 47.50%, with notable improvements across all thresholds, especially at mAP$_{0.75}$ and mAP$_{0.5}$. Overall, the results demonstrate the importance of utilizing both spatial (ROI) and subject-level features for optimal performance in handling confounding variables.

6.4.5.4 Study on Cluster Methods

To verify the effectiveness of different clustering methods within our ACIM, we conduct experiments using six different clustering methods, as shown in Table 6.13. Spectral Clustering (SC) [251] utilizes the eigenvalues of a similarity matrix to reduce dimensionality before applying a clustering algorithm. It is particularly effective for identifying non-convex clusters. Agglomerative Clustering (AC) is a hierarchical method that iteratively merges clusters, represented in a dendrogram [252]. DBSCAN groups closely packed points, marking outliers. Suitable for arbitrary shapes and robust to noise [253]. Gaussian Mixture Model (GMM) assumes data is generated from a mix of Gaussian distributions, using the EM algorithm for parameter estimation [254]. OPTICS extends DBSCAN to handle varying densities and generates a reachability plot for structure visualization [255]. K-means++ improves K-means by optimizing centroid initialization, reducing poor clustering risks [256].

Among the methods, SC achieves moderate performance with an average mAP of 46.53% but shows limited effectiveness at higher thresholds like $mAP_{0.95}$ (25.6%). Similarly, AC and DBSCAN perform comparably, with average mAP scores of 46.17% and 46.13%, respectively, showing slightly weaker results overall. GMM and OPTICS outperform these methods, achieving average mAPs of 46.91% and 47.01%, with OPTICS delivering slightly better performance across all thresholds. Notably, Kmeans++ achieves the highest performance, with an average mAP of 47.50% and the best scores across all thresholds, including $mAP_{0.5}$ (59.16%), $mAP_{0.75}$ (56.77%) and $mAP_{0.95}$ (26.57%). These results demonstrate that Kmeans++ is the most effective clustering method for this task, likely due to its ability to initialize centroids more efficiently, leading to improved convergence and better cluster assignments.

Table 6.13: Ablation study on cluster methods.

Cluster Method	**mAP**$_{0.5}$	**mAP**$_{0.75}$	**mAP**$_{0.95}$	**Avg.**
SC	58.27	55.72	25.6	46.53
AC	57.87	55.38	25.24	46.17
DBSCAN	57.89	55.16	25.34	46.13
GMM	58.51	56.03	26.19	46.91
OPTICS	58.68	56.24	26.12	47.01
Kmeans++	**59.16**	**56.77**	**26.57**	**47.50**

6.4.5.5 Study on Confounder Size

Table 6.14 presents an ablation study examining the effect of different confounder sizes within our ACIM on the model's performance, evaluated through mAP$_{0.5}$, mAP$_{0.75}$, and mAP$_{0.95}$, and the average mAP. Using a confounder size of 64 yields an average mAP of 46.52%, with relatively modest performance across all thresholds. Increasing the confounder size to 128 results in the highest performance, achieving an average mAP of 47.50%, with the best results for mAP$_{0.5}$ (59.16%) and mAP$_{0.75}$ (56.77%). This suggests that a size of 128 offers a balanced representation that captures sufficient context while avoiding redundancy or noise, leading to improved accuracy.

However, as the confounder size continues to grow beyond 128, the performance begins to degrade. For instance, sizes of 256 and 512 result in lower average mAP scores of 46.34% and 46.07%, respectively. This decline may indicate that overly large confounder sizes introduce unnecessary complexity or dilute meaningful features, which negatively impacts the model's ability to focus on relevant information.

Interestingly, while a confounder size of 1024 achieves the highest mAP$_{0.95}$ (26.73%), its

Table 6.14: Ablation study on confounder sizes.

Confounder Size	**mAP**$_{0.5}$	**mAP**$_{0.75}$	**mAP**$_{0.95}$	**Avg.**
64	58.19	55.66	25.7	46.52
128	**59.16**	**56.77**	26.57	**47.50**
256	57.55	55.03	26.44	46.34
512	57.13	54.71	26.37	46.07
1024	57.42	55.23	**26.73**	46.11

overall performance remains suboptimal, with an average mAP of 46.11%. These findings highlight that selecting an appropriate confounder size enables a better representation of the confounders.

6.4.5.6 Study on Pre-trained Dataset

In this ablation study, we compare the performance of models pretrained on three distinct datasets: ImageNet, SSv2, and Kinetics on the task of action detection, as shown in Table 6.15. ImageNet, widely used for image classification tasks, contains millions of labelled images spanning numerous classes and serves as a foundational benchmark for visual tasks [145]. SSv2 (Something-Something V2) is a large-scale video dataset focused on human-object interactions, specifically designed for action recognition tasks [257]. Kinetics, another video dataset, consists of a wide variety of human actions, making it highly suitable for temporal dynamics and action recognition [147].

Pre-training on the ImageNet dataset results in the lowest performance, with an average mAP of 24.72% and particularly weak results across all thresholds, such as mAP$_{0.95}$ (12.58%). This indicates that ImageNet, being designed for image-level tasks, provides limited benefits

Table 6.15: Ablation study on pre-trained datasets.

Pretrain Method	**mAP**$_{0.5}$	**mAP**$_{0.75}$	**mAP**$_{0.95}$	**Avg.**
ImageNet	31.73	29.84	12.58	24.72
SSv2	52.28	50.15	23.06	41.83
Kinetics	**59.16**	**56.77**	**26.57**	**47.50**

for temporal action detection tasks that require capturing temporal dependencies. In contrast, pre-training on the SSv2 dataset yields significantly better results, with an average mAP of 41.83%. The improvement highlights the effectiveness of using a video dataset like SSv2, which contains motion and temporal information, making it more suitable for action detection tasks. Notably, pre-training on the Kinetics dataset achieves the highest performance, with an average mAP of 47.50% and the best scores across all thresholds, including mAP$_{0.5}$ (59.16%), mAP$_{0.75}$ (56.77%) and mAP$_{0.95}$ (26.57%). The superior results suggest that Kinetics, with its diverse and large-scale action-focused data, provides a more comprehensive representation of temporal features, thereby significantly enhancing the model's ability to detect motions. These findings underline the importance of selecting a pre-training dataset that aligns closely with the target task's requirements.

6.4.6 Generalization Analysis

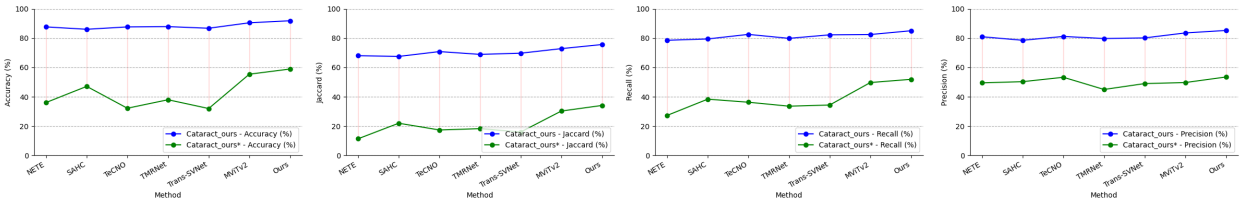
Table 6.16 presents a comprehensive analysis of phase recognition generalization experiments between the CATARACTS and Cataract_ours datasets. The evaluation is conducted at both video-level (accuracy) and phase-level (precision, recall, and Jaccard). Our method demonstrates superior performance in both transfer directions, Cataract_ours → CATARACTS and CATARACTS → Cataract_ours, with consistently high accuracy and phase-level metrics. The superior results indicate that our model effectively captures the underlying causal

Table 6.16: Comparison of generalization experiments between the Cataract_ours and CATARACTS.

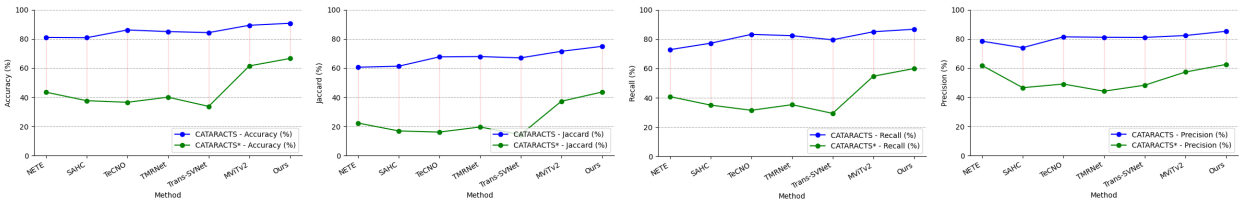
Method	**Video-level**	**Phase-level**		
	Accuracy	**Precision**	**Recall**	**Jaccard**
Cataract_ours → CATARACTS				
NETE	43.5±7.9	61.8±28.7	40.8±33.6	22.4±16.0
SAHC	37.7±6.1	46.2±25.5	31.5±34.2	12.4±12.1
TeCNO	36.7±7.6	49.1±28.7	31.0±34.5	16.2±14.1
TMRNet	40.1±8.6	44.2±24.2	35.1±32.4	17.5±13.8
Trans-SVNet	38.3±7.2	43.1±28.4	31.2±34.9	13.3±12.1
MViTv2	55.6±5.9	57.3±20.2	34.7±25.5	37.3±16.1
Ours	**66.6±6.8**	**62.5±16.7**	**59.9±20.6**	**43.6±13.9**
CATARACTS → Cataract_ours				
NETE	36.1±13.4	49.5±30.8	27.3±36.2	11.4±11.0
SAHC	47.1±14.5	50.3±30.0	39.0±24.8	22.0±16.5
TeCNO	41.6±13.0	50.8±33.2	35.1±27.0	22.4±14.1
TMRNet	38.0±16.9	45.3±30.1	34.4±21.9	15.8±13.7
Trans-SVNet	32.0±12.2	50.9±34.3	34.4±21.9	14.0±13.1
MViTv2	55.4±15.7	53.2±21.7	45.9±16.9	32.1±14.9
Ours	**58.9±13.7**	**53.5±24.1**	**51.9±16.8**	**34.1±13.9**

relationships, enabling better generalization across different datasets.

In the Cataract_ours → CATARACTS transfer, our approach achieves the highest video-level accuracy of 66.6±6.8%, significantly outperforming all other methods, including MViTv2,



(a) CATARACTS ->Cataract_ours



(b) Cataract_ours ->CATARACTS

Figure 6.8: Visualization of generalization experiments between Cataract_ours and CATARACTS on accuracy, precision, recall, and jaccard metrics.

which only reaches 55.6±5.9%. At the phase-level, our method also excels, with a precision of 62.5±16.7%, a recall of 59.9±20.6% and a Jaccard of 43.6±13.9%. Moreover, our model achieves the highest Jaccard score of 43.6±13.9%, which further illustrates its ability to accurately segment and recognize phases, resulting in more precise predictions across different phases. In the reverse transfer, CATARACTS → Cataract_ours, our method maintains its lead, achieving a video-level accuracy of 58.9±13.7%, compared to MViTv2's 55.4±15.7%. Additionally, at the phase-level, our method achieves a precision of 53.5±21.4%, a recall of 51.9±16.8%and a Jaccard score of 34.1±13.9%, again outperforming other approaches. The results highlight the proposed method's superior generalization ability and effectiveness in adapting across datasets. A notable advantage of our method lies in its consistently lower standard deviation across both video- and phase-level metrics. This lower variance indicates more stable and reliable predictions, as our model generalizes effectively across different datasets with less fluctuation compared to competing methods.

As shown in Figure 6.8, we visualize the comparison between two generalization experiments, CATARACTS → Cataract_ours and Cataract_ours → CATARACTS, across differ-

ent methods, and present the comparison results on four metrics. The blue line in the figures represents the performance of different methods on the same dataset, while the green line indicates the performance of different methods on cross-dataset generalization experiments. For example, in Figure 6.8 (a), "Cataract_ours*" refers to the performance of a model trained on CATARACTS, evaluated on the Cataract_ours dataset. The red line highlights the differences in experimental results across various metrics. It is evident that the red line for our method is generally shorter across all metrics, further demonstrating the superior generalization capability of our approach compared to existing methods.

6.4.7 Visualisation and Analysis

6.4.7.1 Action Detection

Figure 6.9 presents the visual results of our method compared to existing approaches in the action detection task. The datasets of this task involve numerous classes, with each instrument corresponding to multiple actions. Since actions performed by the same instrument often have similar spatial features, this task poses significant challenges. As shown in the first row of Figure 6.9 (a), on the Cataract dataset, the samples with the actions "lens-hook pull" and "phacoemulsifier push" are rare in the dataset, making them difficult to identify accurately. While other methods misclassify these actions as other multi-sample classes, such as "lens-hook grasp" and "phacoemulsifier aspirate", our method detects the current action with high confidence and accuracy.

On the GraSP dataset, as shown in the second row of Figure 6.9 (b), one frame contains three instruments and five actions, making it extremely challenging to detect each action fully and accurately. This requires the model to not only identify each instrument accurately but also extract the spatio-temporal features of each instrument to predict the corresponding

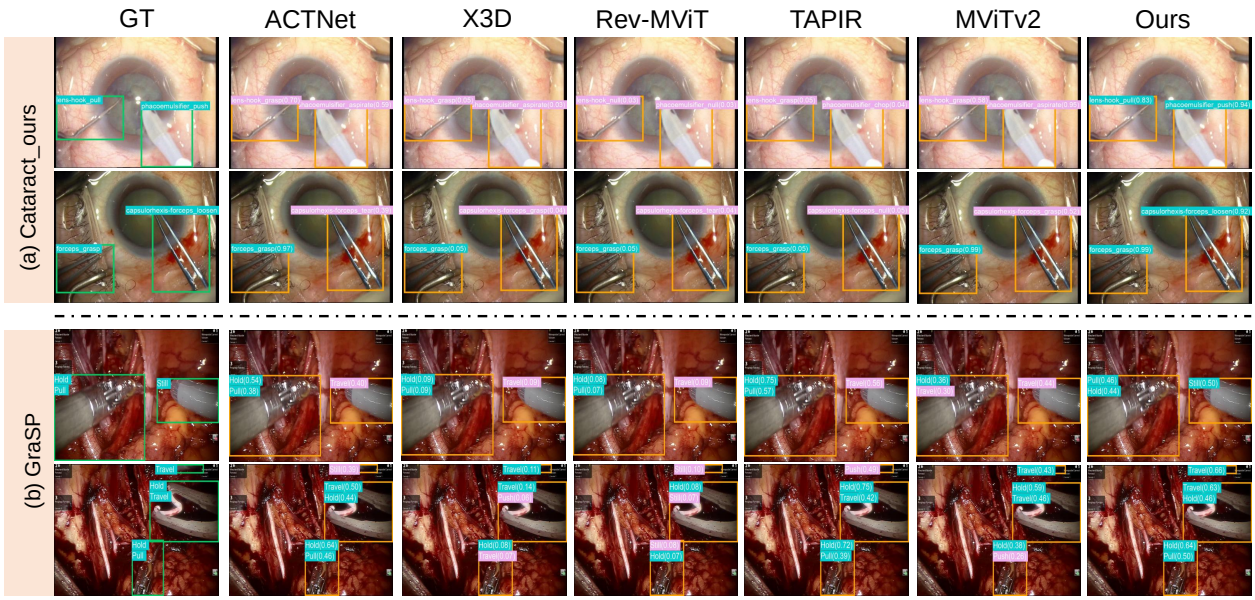


Figure 6.9: Qualitative comparisons for action detection. Action detection locations of Ground Truth (GT) are indicated by green boxes, while the predicted action locations are shown with orange boxes. We also present the recognition results for GT (sky blue), right predictions (sky blue) and wrong predictions (light purple), along with the model's confidence.

actions. While other methods can only partially detect actions, our method is the only one capable of accurately and fully detecting all actions in the current frame, maintaining high confidence levels. These examples further demonstrate the effectiveness and robustness of our method in action detection while also highlighting its advantage in detecting few-shot classes.

6.4.7.2 Step Recognition

We illustrate the visual results of step recognition on full video sequences in Figure 6.10. We select video clips that cover most surgical steps and present the typical workflow procedures in both laparoscopy and cataract surgery. Figure 6.10 shows the recognition results using

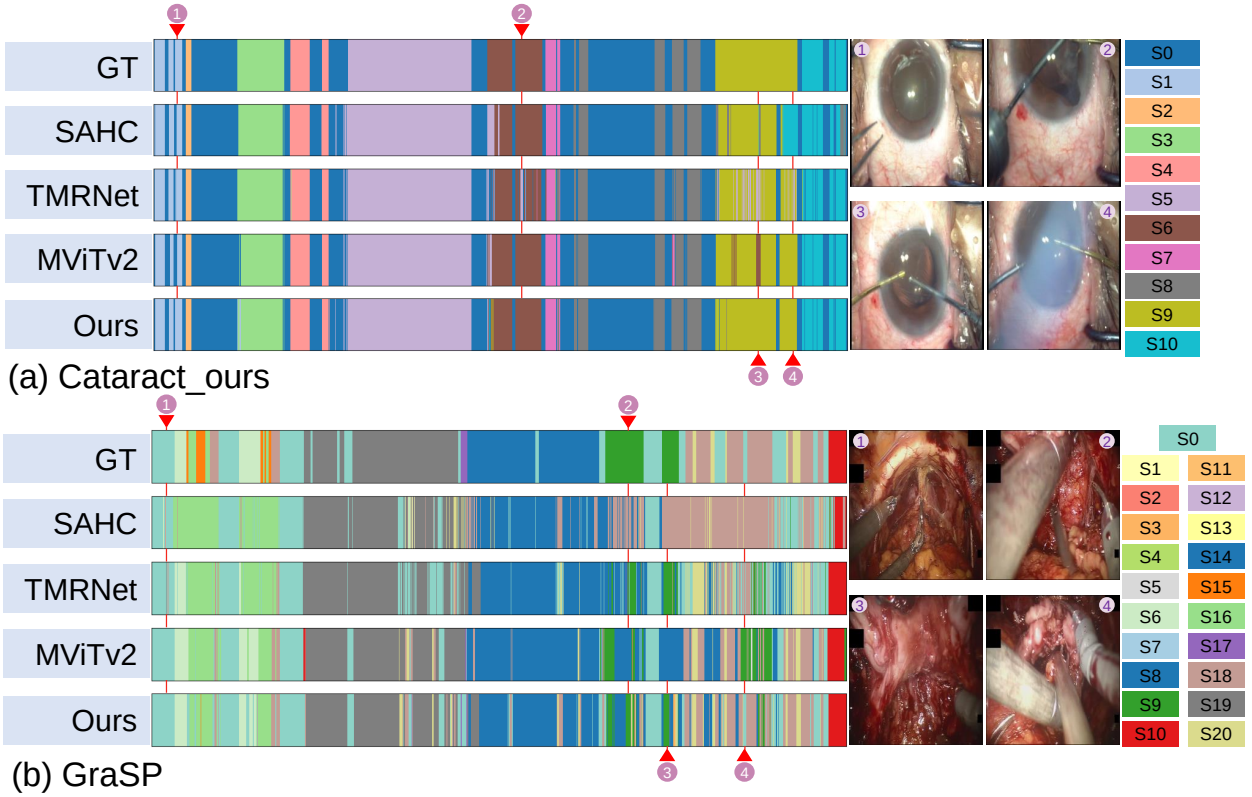


Figure 6.10: Color-coded ribbon illustration for two complete surgical videos on step recognition. The time axes are scaled for better visualization.

color-coded ribbons across five different methods from two datasets: Cataract_ours and GraSP. Additionally, we visualize segments that are prone to misclassification, as indicated by the numbered ribbons in the figure.

On the Cataract_ours dataset, it is evident that our method's class distribution is most similar to the GT. For instance, at position 3, the recognition of class S9 is challenging due to the incomplete instrument visibility in the video. While other methods often misclassify this step as S6 or S5, our method accurately identifies the correct step by leveraging multifaceted information. On the GraSP dataset, there are 21 step classes, and the complexity of endoscopic operations further increases the difficulty of recognition. For example, at position 3, the instrument occupies a small portion of the frame. At positions 2 and 4, the

instruments are partially obscured, making it difficult to detect them fully. In these cases, the recognition task heavily relies on the model's ability to read and analyze spatio-temporal context. Despite these challenges, our method consistently captures key, relevant information and correctly identifies the steps, outperforming existing methods. This demonstrates the effectiveness and robustness of our approach in step recognition tasks.

6.4.7.3 Phase Recognition

We present the visual results of phase recognition on full video sequences in Figure 6.11. Similar to step recognition, Figure 6.11 shows the recognition results using color-coded ribbons across five different methods from two datasets: Cataract_ours and GraSP. On the Cataract_ours dataset, at positions 1 and 2, due to instrument blurring, existing methods misclassify these phases as P0, while our method accurately identifies them as P1. At position 4, due to the similarity between instruments in phases P3 and P4, other methods incorrectly label this phase as P4, whereas our method correctly classifies it.

On the GraSP dataset, at position 3, the overlap between instruments makes it easy to misclassify this phase as P0. However, our method effectively incorporates causal information relevant to phase recognition, leading to the correct identification of this phase.

6.5 Summary

In this chapter, we propose a causality-inspired framework for surgical video representation learning, providing a unified solution for tasks across multiple granularity levels, including phase recognition, step recognition, and action detection. Specifically, we first utilize cross-attention blocks to build a subject-centric module to make our model focus on instruments' key role in surgical activities, offering a more efficient, compact representation by emphasizing

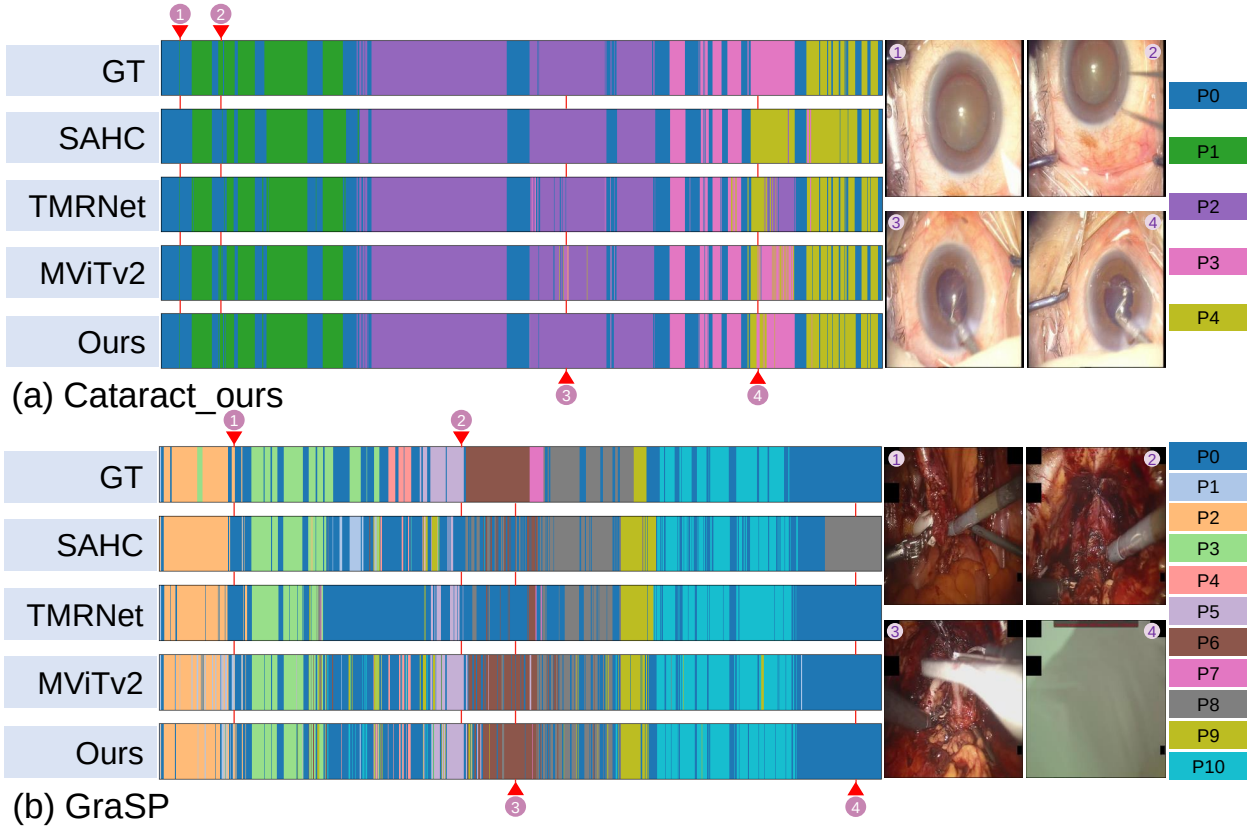


Figure 6.11: Color-coded ribbon illustration for two complete surgical videos on phase recognition. The time axes are scaled for better visualization.

instrument-centric information. By integrating causal intervention techniques, our method effectively addresses the confounding biases that often arise in the understanding of the surgical scene, ensuring that the model learns pure causal relationships rather than spurious correlations. The results demonstrate that our method significantly improves recognition performance while mitigating the impact of confounders. This advancement contributes to the development of more reliable computer-assisted surgical systems, capable of providing context-aware assistance to the surgical team. Future work will explore further refinements of the deconfounding mechanism and its application to a broader range of surgical tasks.

Chapter Seven

A Foundation Model for Multi-level Surgical Scene Understanding

The preceding chapters have built a comprehensive approach to surgical understanding: Chapter 4 focused on fine-grained detection, Chapter 5 built a hierarchical framework, and Chapter 6 introduced a causal framework to enhance robustness and mitigate spurious correlations. However, a persistent challenge highlighted in Chapter 1 remains: the limited generalization of models across different clinical environments (domain shift). The models developed thus far, while robust, have not been validated in a large-scale, multi-center generalization study. This final technical chapter addresses this critical gap by proposing a foundation model specifically designed and validated for robust multi-level workflow recognition across multiple clinical centers.

7.1 Introduction

The significant global burden of ophthalmic diseases has led to a substantial increase in surgical procedures, making them among the most frequently performed in clinical practice [258–260]. The outcomes of these delicate surgeries are critically dependent on procedural precision, which directly impacts patient recovery and visual prognosis. However, the conventional analysis of surgical workflows relies heavily on expert review, which is often subjective and suffers from significant inter-rater variability, leading to inconsistencies and a lack of standardization across different surgeons and institutions [154, 261]. This variability can obscure crucial insights into the factors that determine postoperative success, such as intraoperative decision-making and the sequence of actions. Furthermore, the increasing complexity of modern techniques, including minimally invasive approaches, intensifies the need for robust and objective analytical tools. Recent advances in AI present a promising path toward an objective and scalable analysis of surgical workflows, with significant applications in training, quality assurance, and procedural optimization [6, 262, 263].

While AI has made considerable strides in surgery, much of the research has focused on MIS under endoscopic vision, which benefits from relatively standardized workflows and imaging [47, 264, 265]. Ophthalmic surgery, however, introduces a unique set of challenges. In contrast to MIS, which often uses a more uniform set of tools, ophthalmic procedures like cataract surgery require a diverse array of specialized microsurgical instruments, such as phacoemulsification tips, designed for specific and delicate ocular structures [42]. This instrument diversity, combined with the fine-scale anatomical environment of the eye, presents a sharp contrast to the more generalized framework of endoscopic MIS. Consequently, AI systems developed for MIS cannot be directly transferred to ophthalmology; they require tailored algorithms capable of addressing the unique technical and structural complexity of microsurgery.

Within ophthalmic surgery, AI research has primarily concentrated on high-volume, standardized procedures like cataract surgery, where large datasets have enabled progress in workflow recognition [103, 189, 266]. However, a critical challenge remains that significantly limits the real-world applicability of these models: poor generalization. Models trained on data from one hospital often fail to perform reliably when tested on surgical videos from another medical center due to variations in technique, equipment, and camera setups [47, 267]. This "domain shift" is a well-recognized barrier in medical AI, and overcoming it is essential for developing tools that are robust and trustworthy enough for widespread clinical use [267]. The need to move beyond single-center models toward a more universal solution motivates the development of a foundational AI system capable of understanding cataract surgery across these diverse conditions.

To address these limitations, this thesis argues that a foundation model approach is the most effective strategy for achieving robust domain generalization. Unlike single-center models that overfit to specific environments, a foundation model trained on a large and diverse multi-center corpus (as detailed in Section 7.3.1) is forced to learn domain-invariant repre-

sentations. By learning the fundamental patterns of surgery rather than superficial, hospital-specific cues, such a model can overcome the domain shift barrier. Therefore, we propose the Comprehensive Ophthalmology Surgery (COS) Model, a pioneering foundation model in surgical AI designed for holistic and multi-level workflow understanding in cataract surgery, as illustrated in Figure 7.1. Constructed upon a unified network architecture, the COS Model simultaneously performs hierarchical task recognition—identifying phases, steps, and sub-steps—to enable a detailed decoding of the intraoperative workflow. It demonstrates robust versatility across various cataract surgery contexts without requiring task-specific modifications. Through extensive multi-center validation using surgical videos from operators with diverse expertise across multiple institutions, the model confirms its strong performance, outperforming existing methods for procedural segmentation.

7.2 Methodology

7.2.1 Overview

Our Comprehensive Ophthalmology Surgery (COS) model implements a unified vision-based architecture for surgical workflow recognition across diverse ophthalmic procedures (Figure 7.2). The framework employs a hierarchical feature extraction approach that decouples spatial and temporal representation learning. Initially, robust spatial features are extracted through a Vision Transformer (ViT) encoder initialized with parameters from foundation model pretraining (Figure 7.1), enabling effective capture of anatomical structures, instrument configurations, and tissue manipulation patterns within individual frames. This spatial encoding is followed by dedicated temporal modeling to capture procedural dynamics, technical transitions, and surgeon-specific workflow patterns across sequential frames. This architecture design addresses the fundamental challenge in surgical video analysis—balancing

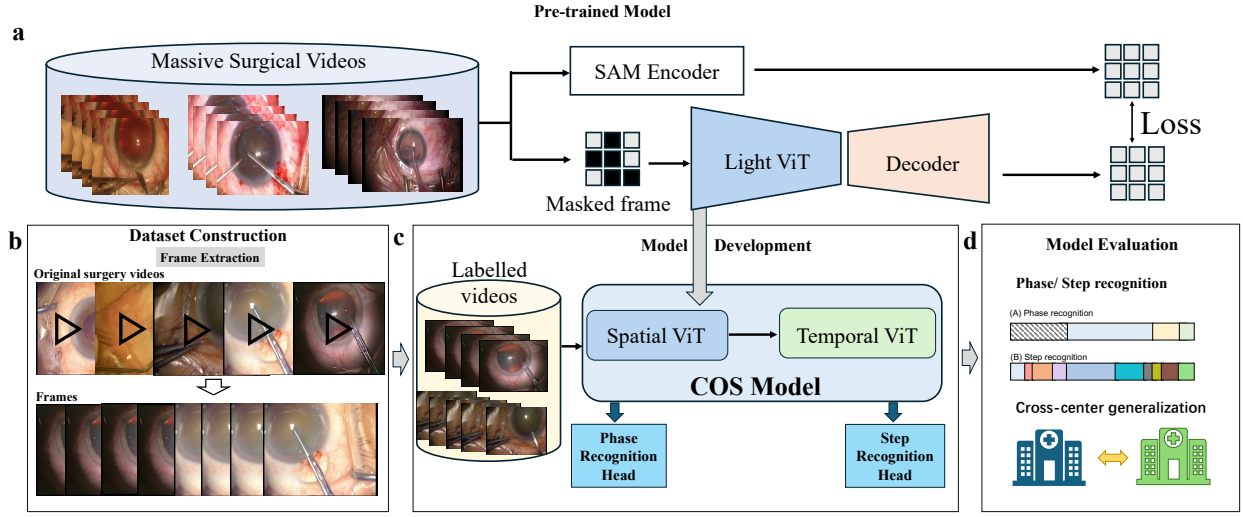


Figure 7.1: Workflow of this study. a. A pre-trained model: massive surgical videos from publicly available datasets and unlabeled videos are collected to pretrain the light Vision Transformer (ViT). b. Dataset Construction: multi-center ophthalmology surgical videos were converted to images for dataset construction. All frames from the datasets were annotated by annotators and reviewed by the expert surgeon. The annotation includes phase and step. c. Model Development: labeled surgical videos are used to train the COS model, which included pretrained ViT capturing spatial information and temporal ViT for temporal features among frames. d. Model evaluation: our COS model is tested on multi-center datasets. Its performance was evaluated using some metrics and feature/results visualization and compared with surgeon. COS Model: Comprehensive Ophthalmology Surgical Model.

fine-grained spatial understanding (critical for instrument and anatomical recognition) with temporal continuity modeling (essential for phase boundary detection and operation smoothness characterization).

7.2.2 Spatial Vision Transformer (ViT)

Recently, the vision transformer [268] has achieved most of the focus for feature learning from images and has become a classical method for image feature representation. Thus, we adopt the vision transformer as the backbone to extract the spatial features of frames in one GOP. To enhance the ability of the feature representation, we adopt the SAM-leveraged Masked Image pertaining (SAMI) framework introduced by Effcient-SAM [269] to pre-train the model on a relatively large set of images without labels. The backbone of the SAM model is adopted as the teacher model, and one ViT-Tiny (ViT_t) model [270] is adopted as the student model. Given one image, we randomly mask some pixels in it, and then the unmasked part is fed to the student model mapped as the encoded features. Then, the unmasked features and the masked tokens will feed to one decoder comprising several cross-attention modules to reconstruct the features that the teacher models with the whole image extracts. Different from the original Effcient-SAM [269], we introduced a random selection of one blur algorithm as an additional augmentation step. Concurrently, ColorJitter was applied as a data augmentation technique. The blur methods incorporated in our study included simple blur, box blur, and Gaussian blur.

7.2.3 Temporal Vision Transformer (ViT)

The temporal information between frames is also a significant part of the tasks, such as video phase classification. As the spatial ViT extracts the spatial features of frames, we can use the class token in the ViT to represent the features of one frame inspired by [268]. Thus, the temporal information can be extracted from the class tokens from frames. Given K frames, we get the spatial features $F_s \in \mathbb{R}^{K \times L \times D}$ after the spatial ViT where K is the number of frames, L is the length of the features which equal $H \times W + 1$, D is the dimension of the features. Then, we can get the class token $F_{cls} \in \mathbb{R}^{K \times D}$ from F_s. Following the self-attention

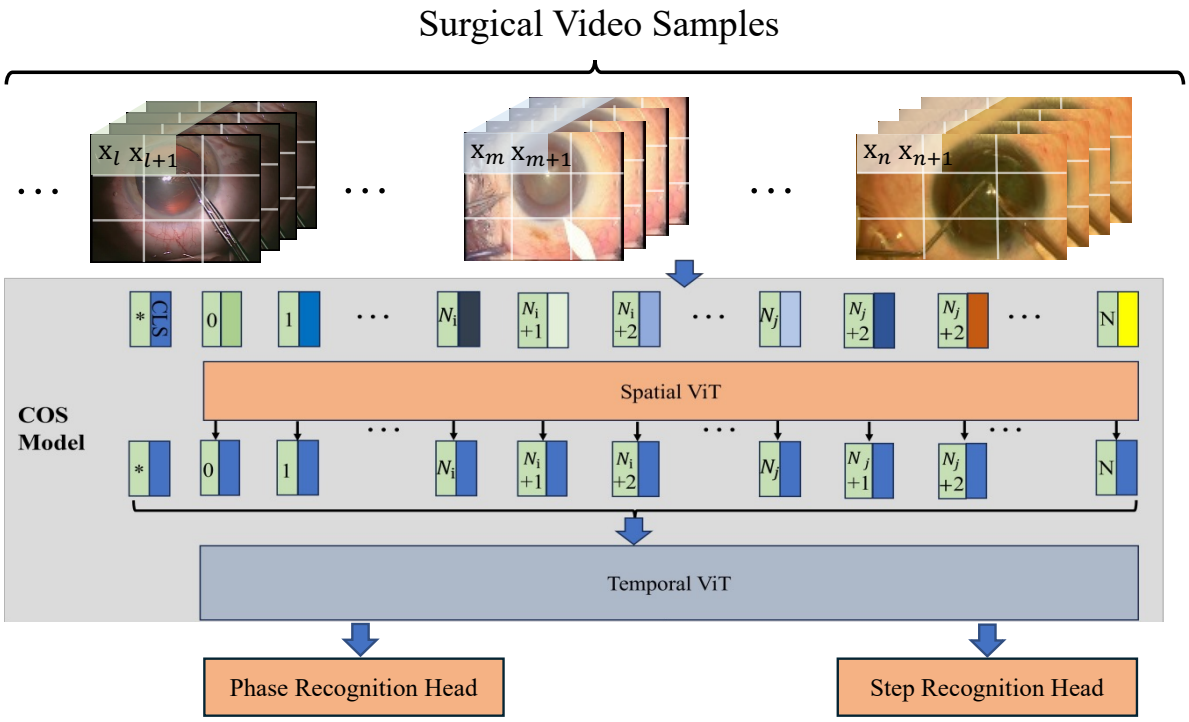


Figure 7.2: Flowchart of our Comprehensive Ophthalmology Surgical (COS) Model for comprehensive ophthalmology surgical learning. The COS Model incorporates a two-step process for surgical learning in ophthalmology. In the first step, spatial features are extracted using the Vision Transformer (ViT) as the encoder, leveraging its pre-trained parameters. Recognizing the unique demands of surgical video analysis, we introduce a second step dedicated to temporal feature learning. This approach ensures a nuanced understanding of both spatial and temporal dimensions crucial for comprehensive surgical education.

mechanism, the query, key, and value are extracted from F_{cls} by the linear project W_q, W_k, and W_v. The attention $att \in \mathbb{R}^{K \times K}$ is generated by the product by query and key, the process is as follows:

$$att = Softmax(W_q(F_{cls}) \cdot W_k(F_{cls})^T) \tag{7.1}$$

Then we can product the value to get the temporal features F_t, which can be described as:

$$F_t = att \cdot W_v(F_{cls}) + F_{cls} \tag{7.2}$$

As the structure of the transformer block, we also use Feed-Forward Network (FFN) to achieve the features in F_t and help with the representation. Thus, the final output is achieved by:

$$F'_t = FFN(F_t) + F_t \tag{7.3}$$

Through this whole process, the interaction between frames is complete and temporal information is captured.

The videoclip classifier in our framework uses MViT [211, 212] to accurately classify the phases and steps of instrument operation. First, the module concatenates contextual features at each time step into a single vector. Then, the vector is processed through average pooling, which averages the combined vectors over the temporal dimension. Lastly, a linear layer maps the pooled spatial-temporal information to phase and step classes.

7.3 Experiments

7.3.1 Cataract Datasets

Ethical approval: All cataract datasets utilized in this study—comprising cataract surgical videos from Shenzhen First People's Hospital and Eye Hospital of Wenzhou Medical University—were collected under institutional review board approvals (reference numbers: LL-KY-2020220, 2023-J-065) with appropriate informed consent procedures in accordance with the Declaration of Helsinki. Comprehensive de-identification protocols were implemented prior to any computational processing or model development activities to ensure patient privacy and data security compliance. All personally identifiable information, including but not limited to patient demographics, case identifiers, and timestamp metadata, was removed following institutional standards for anonymization of surgical video content.

The complete de-identification workflow was independently verified by institutional data protection officers at each contributing center to confirm adherence to both local regulatory requirements and international data protection standards.

Surgery illustration: Cataract surgery represents a high-precision ophthalmic intervention performed when lens opacification significantly impairs visual function and quality of life. The procedural workflow involves creating a microincision, removing the clouded crystalline lens, and implanting an artificial Intraocular Lens (IOL). The technical execution of this procedure directly influences both safety outcomes and visual rehabilitation. Suboptimal surgical technique can precipitate significant complications including endophthalmitis, persistent corneal edema, cystoid macular edema, retinal detachment, or posterior capsular opacification–each potentially compromising visual recovery and necessitating additional interventions. Beyond these complications, technical precision during capsulorrhexis creation, nuclear disassembly, and IOL positioning critically determines postoperative refractive outcomes. Even minor deviations in IOL centration, tilt, or axial position can induce unwanted astigmatism, higher-order aberrations, or refractive surprise that compromises uncorrected visual acuity. Cataract surgery shares fundamental requirements for meticulous tissue handling, precise spatial navigation within confined anatomical spaces, and surgeon-specific decision-making patterns that adapt to individual patient characteristics. These similar technical and cognitive demands make both procedures excellent candidates for computational workflow analysis.

Surgical phase and step Annotations: For computational analysis of cataract surgical workflows, we compiled a comprehensive multi-institutional dataset comprising 182 videos (135,396 frames) from routine phacoemulsification procedures without significant intraoperative complications. This dataset integrated 21 videos (15,771 frames) from Eye Hospital of Wenzhou Medical University (WZH), 50 videos (68,097 frames) from Shenzhen First People's Hospital (SZH), and 101 videos (50,528 frames) from the publicly available Cataract-101

benchmark dataset. All surgical recordings underwent standardized retrospective annotation [271] performed by a multidisciplinary team of ophthalmologists with expertise in anterior segment surgery. The annotation framework employed a hierarchical taxonomy based on the standardized operative protocol established at Klinikum Klagenfurt, with procedural segmentation into 5 distinct phases and 11 constituent technical steps (detailed in Table 7.1). We explicitly incorporated 'idle' segments as a distinct procedural element to enable comprehensive workflow characterization.

Multi-center Datasets: For computational workflow analysis of cataract surgery, we assembled a multi-institutional dataset of 172 phacoemulsification procedures from three distinct sources: Eye Hospital of Wenzhou Medical University (WZH; n=21), Shenzhen First People's Hospital (SZH; n=50), and the publicly available Cataract-101 benchmark collection (n=101) [271]. This diverse dataset enabled rigorous evaluation of our model's generalizability across varying institutional settings. While phacoemulsification technique follows relatively standardized procedural steps across institutions, we documented notable inter-center variations in instrumentation preferences, phacoemulsification platforms, and surgical adjuncts (Figure 7.3). These technical variations provided an ideal test bed for assessing the robustness of our computational workflow recognition algorithms against real-world procedural heterogeneity. The SZH dataset comprised 50 videos (85,659 frames) acquired at standardized temporal sampling (1 frame per second) and spatial resolution (720 $\times$ 576 pixels). For model development and internal validation, we implemented a stratified random allocation with 80% (40 videos) assigned to the training cohort and 20% (10 videos) to the internal validation cohort.

The WZH dataset (21 videos; 15,771 frames) maintained identical temporal sampling and spatial resolution parameters as the SZH collection but was reserved exclusively for external validation testing. This methodological approach–training exclusively on SZH data and testing on the independent WZH dataset–provided a rigorous assessment of the model's perfor-

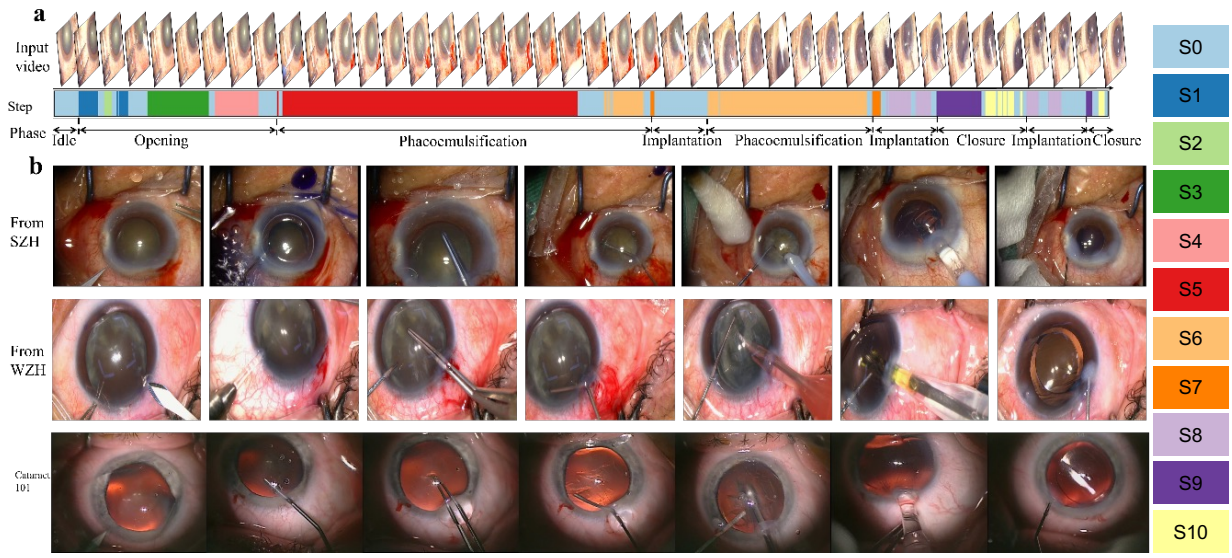


Figure 7.3: The phase and step of a cataract surgery video are sampled in the figure. The instruments and image contrast are different from each center data.

mance when confronted with previously unseen institutional variations in surgical technique, equipment configuration, and procedural workflow.

We further incorporated the Cataract-101 dataset [271], a publicly available resource comprising 101 phacoemulsification procedures (50,528 frames) performed by four surgeons over a nine-month period. While this dataset originally featured step-level annotations (incision, viscoelastic injection, capsulorrhexis, hydrodissection, phacoemulsification, irrigation/aspiration, capsule polishing, IOL insertion, viscoelastic removal), we supplemented these with standardized phase categorizations (opening, phacoemulsification, lens implantation, closure) under the supervision of board-certified anterior segment specialists. These enhanced annotations will be publicly released to support comparative methodological research. The Cataract-101 dataset maintained consistent technical specifications with our institutional collections, featuring standardized temporal sampling (1 fps) and spatial resolution (720×576 pixels). For model validation using this dataset, we implemented a stratified random allocation with 80% (81 videos) designated for training and 20% (20 videos) for

Table 7.1: The details of cataract surgery tasks and generalization to different hospitals.

Hospital	Videos (Frames)	Task
WZH	21 (15771)	Phase, Step
SZH	50 (68097)	Phase, Step
Cataract-101	101 (50528)	Phase, Step
Phase details	P0: Idle, P1: Opening, P2: Phacoemulsification, P3: Implantation, P4: Closure	
Step details	S0: Idle, S1: Incision, S2: Viscous agent injector, S3: Rhexis, S4: Hydrodissection, S5: Phacoemulsification, S6: Irrigation and aspiration, S7: Capsule polishing, S8: Lens implant, S9: Viscous agent removal, S10: Wound hydration	

testing.

This tripartite dataset structure–encompassing SZH (primary training cohort), Cataract-101 (mixed-institutional validation), and WZH (external validation)–enabled comprehensive assessment of our computational workflow recognition model across increasingly challenging generalizability scenarios: from within-institution validation to cross-institutional deployment and finally to entirely unseen surgical environments. This rigorous validation framework addresses a critical limitation in existing surgical video analysis literature, where models often demonstrate degraded performance when applied outside their training domain.

7.3.2 Implementation Details

The model is developed using Pytorch [182] and trained on an NVIDIA RTX A6000 GPU. The training strategy followed the "Model Development" step (c) outlined in Figure 7.1:

the Spatial Vision Transformer (ViT_t), described in Section 7.2.2, was initialized with the parameters from our SAMI pre-training. The new Temporal Vision Transformer (Section 7.2.3) and the final classification head were initialized from scratch.

During the main training phase, all parameters of the model (both the pre-trained spatial encoder and the new temporal/classifier layers) were fine-tuned jointly; no parameters were frozen. We utilize the AdamW optimizer with an initial learning rate of 1e−3 and a weight decay of 0.05. A cosine learning rate scheduler is employed, with a linear warm-up period of 10 epochs, starting at a learning rate of 1e−5. Data augmentation techniques, including 224 × 224 cropping, jittering, and flipping, are applied to enhance the training dataset. Following the approach of [211], the sequence length is set to 16 frames.

Pretrain datasets: For efficient model initialization, we implemented the SAMI framework to distill knowledge from foundation vision models into our surgical workflow analysis architecture. This approach utilizes self-supervised distillation, enabling lightweight derivatives that maintain representation power while achieving computational efficiency that is suitable for clinical deployment. Our pretraining corpus encompassed diverse surgical specialities from multiple institutions: 50 cataract videos (33,055 frames) from the Cataract2020 semantic segmentation benchmark [272] supplemented by 20 additional procedures (329,931 frames) from a collaborating center. All pretraining videos maintained their native temporal resolution to preserve fine-grained motion dynamics critical for surgical technique analysis. Importantly, we ensured no overlap between the pretraining corpus and subsequent task-specific training or validation cohorts to prevent data contamination and enable unbiased performance assessment.

7.3.3 Evaluation Metrics

For the tasks of phase and step recognition, we list the Receiver Operating Characteristic (ROC) curve in the experiments. The performance is evaluated using two commonly utilized metrics: Area Under the ROC Curve (AUC) and specificity. When the decision threshold is varied and a number of points on the ROC curve $[P(Fp) = \alpha, P(Tp) = 1 - \beta]$ have been obtained, the simplest way to calculate the area under the ROC curve is to use trapezoidal integration [273],

$$\mathrm{AUC} = \sum_i \left\{ (1 - \beta_i \cdot \Delta\alpha) + \frac{1}{2}[\Delta(1-\beta) \cdot \Delta\alpha] \right\}, \tag{7.4}$$

where

$$\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1}), \Delta\alpha = \alpha_i - \alpha_{i-1} \tag{7.5}$$

The specificity is derived from True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), defined as:

$$specificity = \frac{TN}{TN + FP} \tag{7.6}$$

7.3.4 Generalization Experiments

To investigate the COS Model's capacity to address the technical diversity of ophthalmic procedures, we evaluated the model across three dimensions: intra-domain performance on core datasets, cross-centre generalizability using external cohorts and comparative analysis against established benchmarks.

7.3.4.1 Quantitative Evaluation

Our model was trained on two independent datasets (Cataract SZH and public Cataract101) and subsequently validated on the external Cataract WZH dataset (Table 7.2-7.5). For sur-

Table 7.2: Performance comparison: Train SZH → Test SZH (avg ± std) across all of the five folds.

	Phase		Step	
Model	AUC	Specificity	AUC	Specificity
COS model	0.9806 ± 0.006	0.9691 ± 0.006	0.9894 ± 0.005	0.9874 ± 0.002
COS model, W/o T	0.9771 ± 0.008	0.9665 ± 0.006	0.9846 ± 0.007	0.9845 ± 0.003
ViT_t, W T	0.9220 ± 0.010	0.9282 ± 0.006	0.9385 ± 0.010	0.9664 ± 0.003
ViT_t, W/o T	0.8992 ± 0.013	0.9223 ± 0.006	0.9183 ± 0.016	0.9641 ± 0.002

gical phase recognition, the COS architecture demonstrated a superior AUC and specificity compared to the baseline vision transformer (ViT_t) across both internal and external validation cohorts (Table 7.2-7.5, Figures 7.4a-d, Figures 7.5-7.8).

Table 7.3: Performance comparison: Train SZH → Test WZH (avg ± std) across all of the five folds.

	Phase		Step	
Model	AUC	Specificity	AUC	Specificity
COS model	0.8110 ± 0.014	0.8727 ± 0.018	0.8775 ± 0.034	0.9405 ± 0.006
COS model, W/o T	0.8730 ± 0.016	0.9097 ± 0.011	0.8617 ± 0.008	0.9362 ± 0.003
ViT_t, W T	0.5782 ± 0.030	0.8046 ± 0.006	0.5808 ± 0.025	0.9112 ± 0.002
ViT_t, W/o T	0.6797 ± 0.023	0.8385 ± 0.006	0.7275 ± 0.018	0.9231 ± 0.002

Remarkably, our COS architecture accurately recognized procedural step sequences without explicit training on interstep relationships (Figures 7.4a-d). The model also demonstrated inherent anomaly detection capabilities, identifying a 129-second aberration during the lens-implantation phase (Figure 7.4g). Subsequent analysis revealed this interval repre-

Table 7.4: Performance comparison: Train Cataract101 → Test Cataract101 (avg ± std) across all of the five folds.

	Phase		Step	
Model	AUC	Specificity	AUC	Specificity
COS model	0.9849 ± 0.005	0.9801 ± 0.004	0.9894 ± 0.002	0.9882 ± 0.002
COS model, W/o T	0.9846 ± 0.003	0.9760 ± 0.002	0.9867 ± 0.002	0.9855 ± 0.001
ViT_t, W T	0.9520 ± 0.01	0.9748 ± 0.007	0.9539 ± 0.008	0.9709 ± 0.002
ViT_t, W/o T	0.9300 ± 0.012	0.9327 ± 0.004	0.9358 ± 0.007	0.9620 ± 0.002

sented an unplanned lens reimplantation following a period of surgical inactivity [274, 275]. Figure 7.3a also illustrates this point. The video in Figure 7.3a was special in that the doctor performed phacoemulsification twice, implantation three times, and closure twice, which could be considered abnormal operations. The predicted results demonstrated that our method could effectively and accurately identify such abnormalities, providing a reference for doctors. This finding underscores the model's capacity to detect deviations from standard surgical workflow, suggesting potential clinical utility for intraoperative anomaly identification.

7.3.4.2 Ablation Study

Ablation studies (Table 7.2-7.5) evaluated the impact of temporal information on algorithmic performance relative to baseline Vision Transformer (ViT_t) models (Figures 7.4e-f, Figures 7.5-7.8). In homogeneous validation scenarios (SZH → SZH, Cataract101 → Cataract101), our complete COS architecture with temporal features consistently demonstrated superior AUC and specificity for both phase and step recognition. For instance, phase recognition in the SZH dataset yielded an AUC of 0.9806 with temporal features versus 0.9220 without.

Table 7.5: Performance comparison: Train Cataract101 → Test WZH (avg ± std) across all of the five folds.

	Phase		Step	
Model	AUC	Specificity	AUC	Specificity
COS model	0.8647 ± 0.038	0.9102 ± 0.019	0.8780 ± 0.043	0.9489 ± 0.005
COS model, W/o T	0.8945 ± 0.017	0.9258 ± 0.011	0.8589 ± 0.024	0.9358 ± 0.006
ViT_t, W T	0.6467 ± 0.032	0.8251 ± 0.008	0.6040 ± 0.006	0.9149 ± 0.002
ViT_t, W/o T	0.7094 ± 0.020	0.8466 ± 0.008	0.7130 ± 0.015	0.9197 ± 0.003

Intriguingly, heterogeneous validation (SZH → WZH, Cataract101 → WZH) revealed that temporal feature removal occasionally improved generalization. This effect was particularly pronounced for the task of phase recognition (AUC 0.8730 without temporal features vs. 0.8110 with temporal features in SZH → WZH testing). This suggests temporal information enhances pattern recognition within consistent surgical environments but may introduce dataset-specific biases that impede cross-center generalization. These findings advocate for context-dependent temporal feature implementation, balancing intra-dataset performance with inter-dataset transferability.

These two experiments (phase and step recognition) primarily validate three key findings: First, the incorporation of temporal information into our COS model positively impacts the results, effectively improving recognition accuracy. Second, our approach demonstrates generalizability across diverse datasets. Despite variations among the datasets, the experimental outcomes confirm that our method possesses strong generalization capabilities. Third, the recognition difficulty varies across different operational stages or steps, resulting in performance fluctuations. For example, the "irrigation-and-aspiration" step is more susceptible to confusion with other steps, leading to comparatively lower recognition accuracy.

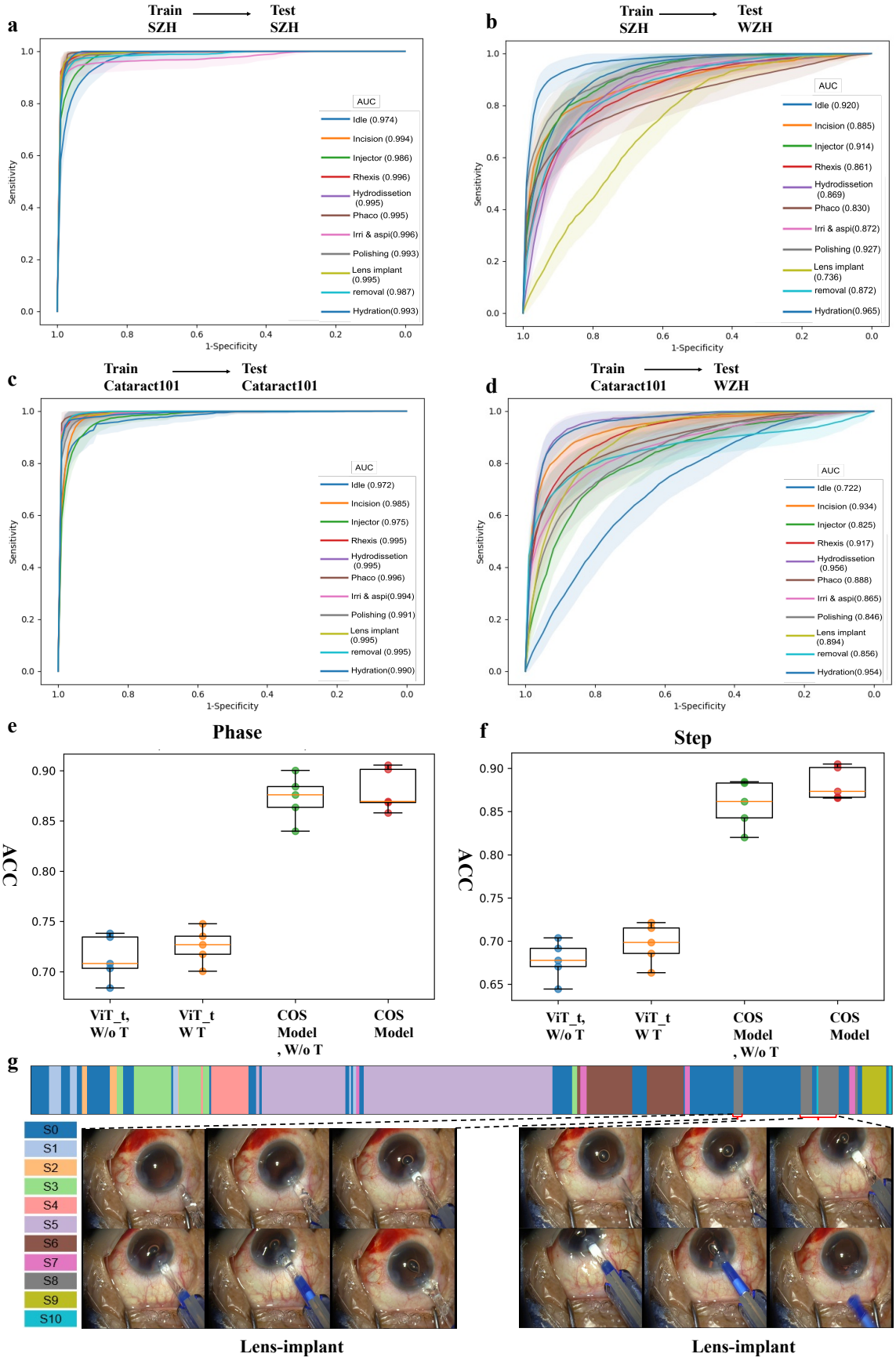

a
Train SZH → Test SZH
Sensitivity
1-Specificity
AUC
Idle (0.974)
Incision (0.994)
Injector (0.986)
Rhexis (0.996)
Hydrodissetion (0.995)
Phaco (0.995)
Irri & aspi(0.996)
Polishing (0.993)
Lens implant (0.995)
removal (0.987)
Hydration(0.993)
b
Train SZH → Test WZH
Idle (0.920)
Incision (0.885)
Injector (0.914)
Rhexis (0.861)
Hydrodissetion (0.869)
Phaco (0.830)
Irri & aspi(0.872)
Polishing (0.927)
Lens implant (0.736)
removal (0.872)
Hydration(0.965)
c
Train Cataract101 → Test Cataract101
Idle (0.972)
Incision (0.985)
Injector (0.975)
Rhexis (0.995)
Hydrodissetion (0.995)
Phaco (0.996)
Irri & aspi(0.994)
Polishing (0.991)
Lens implant (0.995)
removal (0.995)
Hydration(0.990)
d
Train Cataract101 → Test WZH
Idle (0.722)
Incision (0.934)
Injector (0.825)
Rhexis (0.917)
Hydrodissetion (0.956)
Phaco (0.888)
Irri & aspi(0.865)
Polishing (0.846)
Lens implant (0.894)
removal (0.856)
Hydration(0.954)
e
Phase
ACC
ViT_t, W/o T
ViT_t W T
COS Model , W/o T
COS Model
f
Step
g
S0
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
Lens-implant
Lens-implant

Figure 7.4: Step/phase recognition for Cataract surgery. a. and b. The ROC was trained on the SZH dataset and tested on the SZH and WZH datasets, c. and d. The ROC trained on the publicly available Cataract101 dataset and tested on Cataract101 and WZH datasets. e. and f. The ablation study results prove the effectiveness of each component by phase and step recognition. g. The colour-coded ribbon illustration for a complete surgical video on step recognition. The COS model discovered outlier behaviors of two-times lens implant, which can indicates wrong operations, or complicated surgery.

7.4 Summary

This study presents the Comprehensive Ophthalmology Surgery (COS) Model, a foundation model for multi-level surgical workflow understanding in cataract surgery. The research addresses the critical challenge of poor cross-center generalization in existing cataract surgical AI systems due to variations in technique, equipment, and imaging setups.

The COS model employs a unified architecture combining Vision Transformer (ViT) for spatial feature extraction with temporal modeling to capture procedural dynamics. Through multi-center validation using datasets from three institutions (WZH, SZH, and Cataract-101), the model demonstrates significantly improved performance compared to the baseline ViT_t architectures (as detailed in Tables 7.2 and 7.4), with AUC values exceeding 0.98 for both phase and step recognition in homogeneous settings. Importantly, the model exhibits robust cross-center generalization capabilities while maintaining effective performance on external datasets.

Additionally, it demonstrates inherent anomaly detection abilities, successfully identifying surgical deviations such as unplanned lens reimplantation during procedures. However, ablation studies reveal that temporal features, while enhancing intra-dataset performance, may introduce dataset-specific biases that occasionally limit cross-center transferability. This

suggests the need for context-dependent implementation strategies. The COS model represents a significant advancement toward developing universal surgical workflow analysis tools capable of operating reliably across diverse clinical environments. These findings support the potential for improved surgical training, quality assurance, and procedural optimization through foundation model approaches in surgical AI.

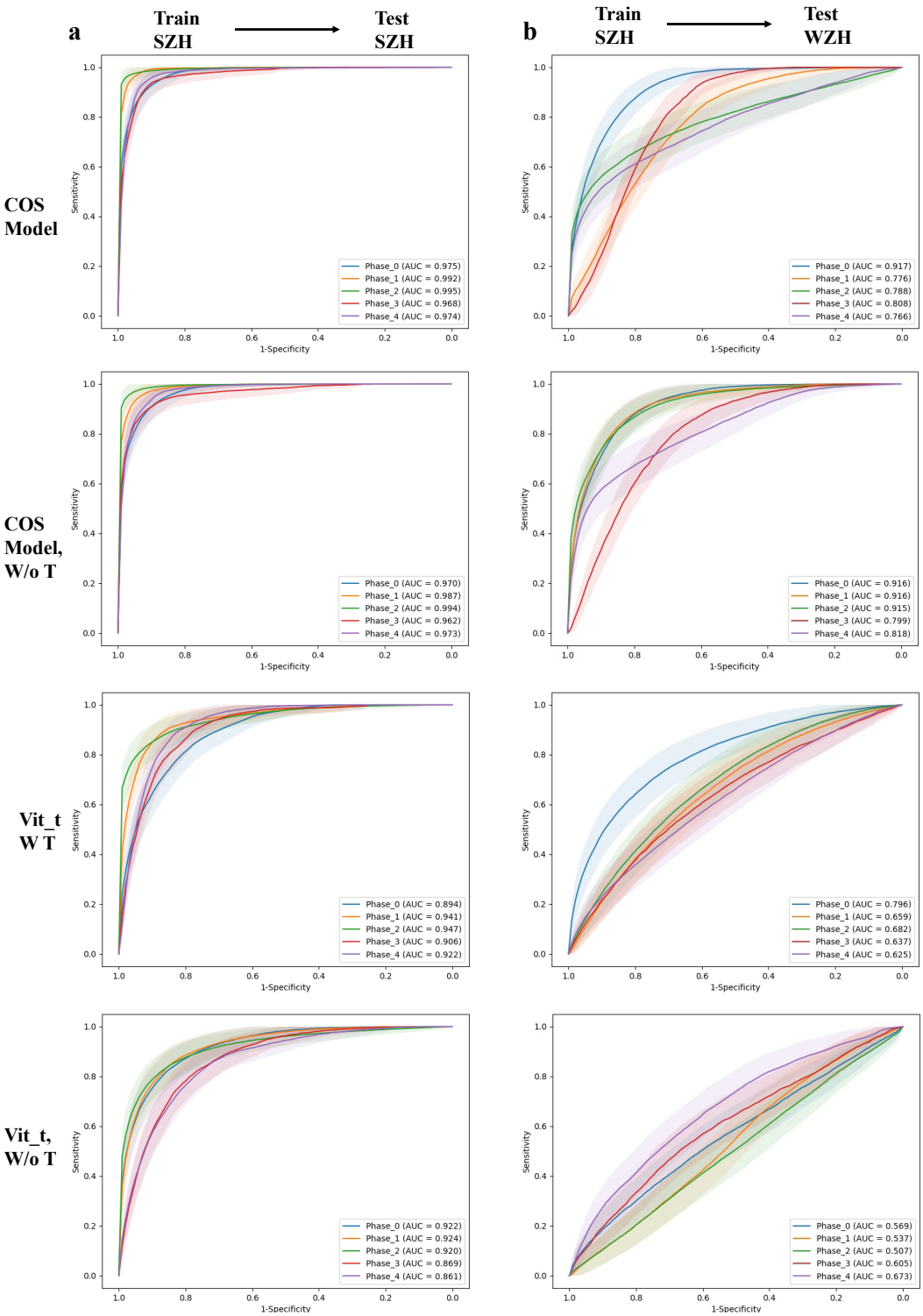


Figure 7.5: The ROC curves of phase recognition by different algorithms for cataract surgery. a. and b. are trained on SZH and test on SZH and WZH datasets.

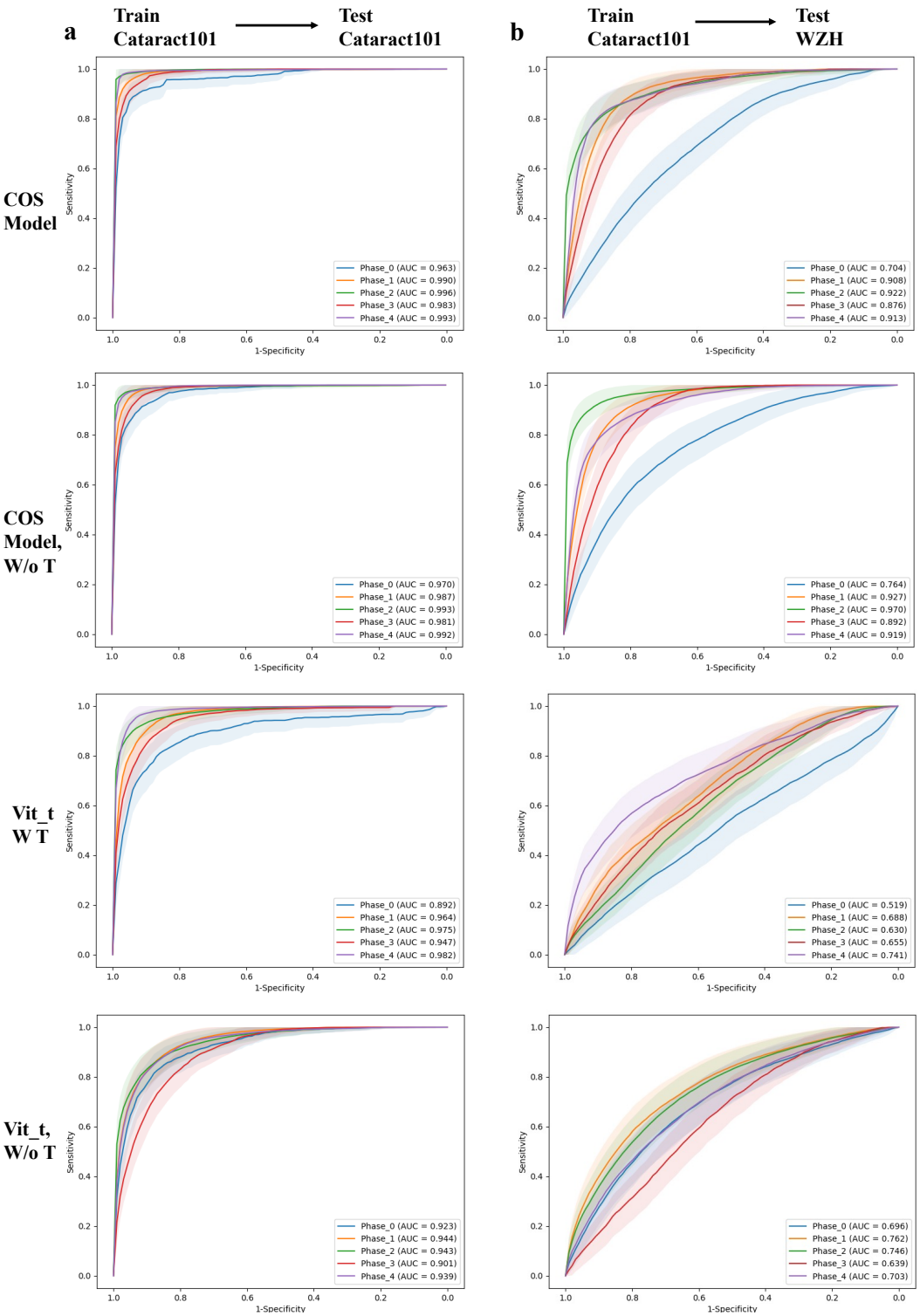


Figure 7.6: The ROC curves of phase recognition by different algorithms for cataract surgery. a. and b. are trained on Cataract101 and test on Cataract101 and WZH datasets.

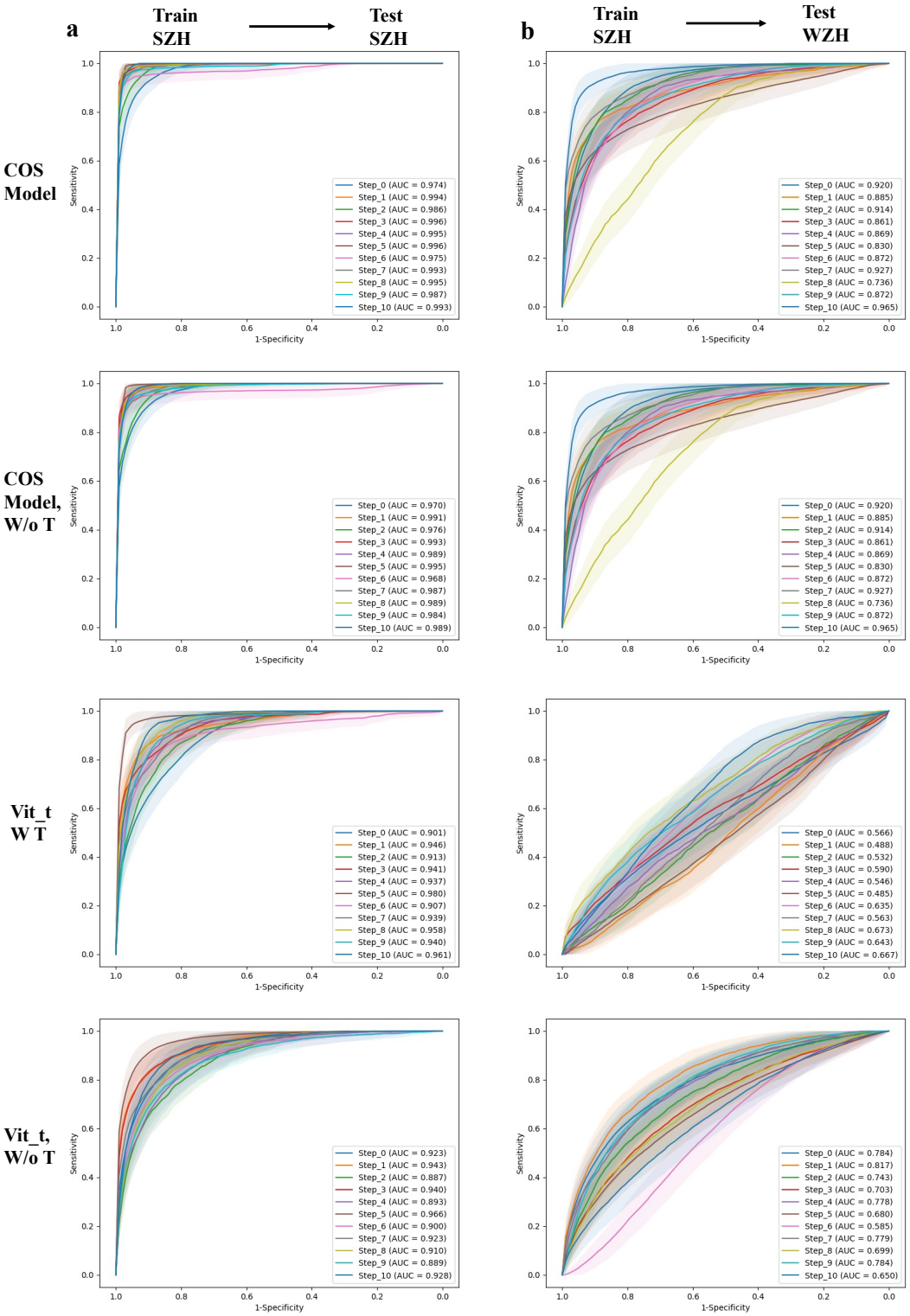


Figure 7.7: The ROC curves of step recognition by different algorithms for cataract surgery. a. and b. are trained on SZH and test on SZH and WZH datasets.

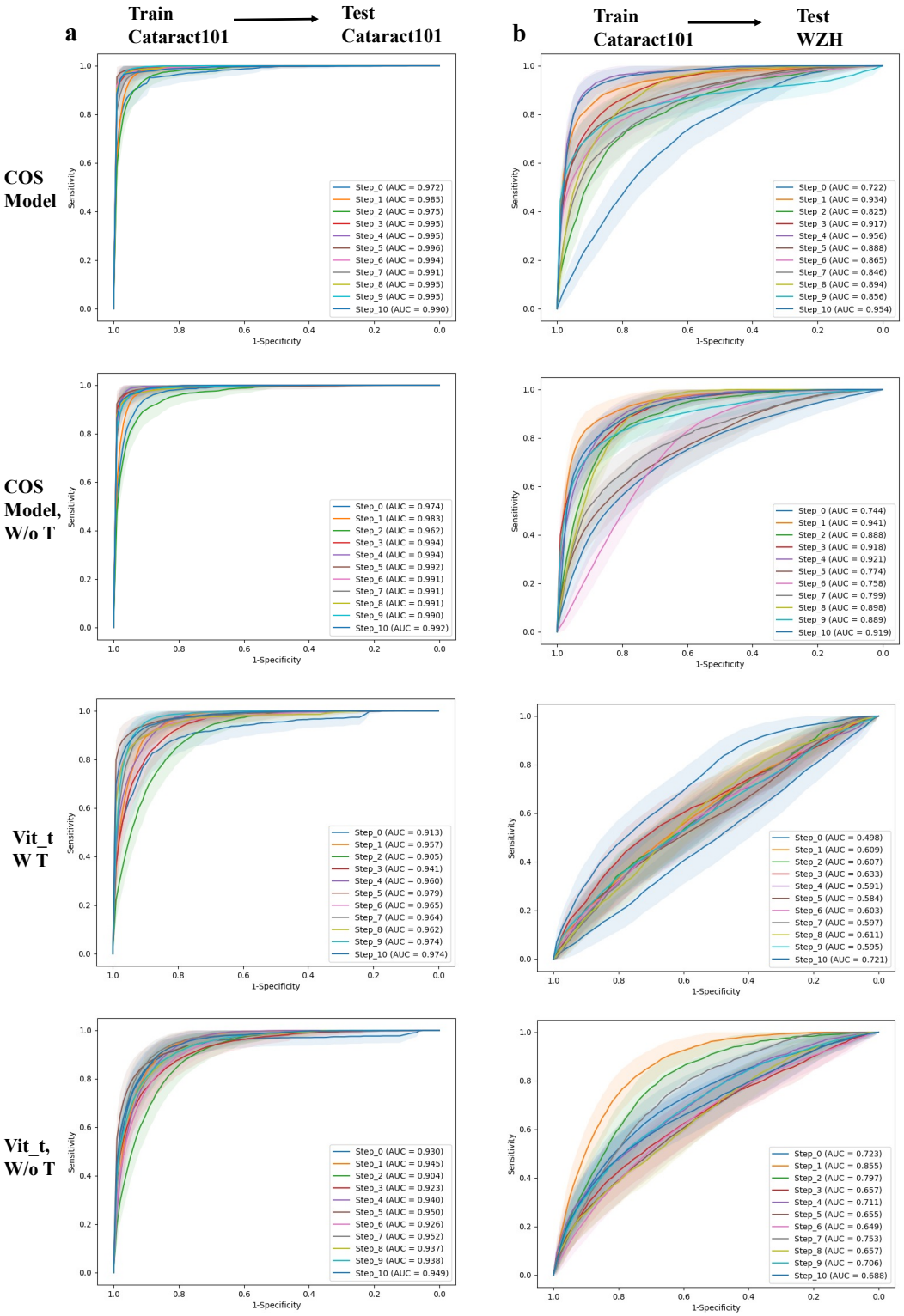


Figure 7.8: The ROC curves of step recognition by different algorithms for cataract surgery. a. and b. are trained on Cataract101 and test on Cataract101 and WZH datasets.

Chapter Eight

Conclusion and Future Work

This chapter presents a summary of the works presented in this thesis. I tackle different objectives related to multi-level surgical scene understanding and highlight the key takeaways of these studies. I conclude this chapter by suggesting perspectives on future research.

8.1 Summary and Conclusion

The primary goal of developing next-generation surgical interventions is to ensure safe and efficient surgery that minimizes pain, allows for fast recovery, and reduces complications. Advances in recent decades have introduced high-tech surgical systems for less invasive and more effective techniques, which has in turn increased the complexity within the operating room. This demands the development of advanced systems to support clinicians across the surgical workflow by inspecting the vast array of data sources that can be recorded. To enable context-aware assistance in the Operating Room (OR), this thesis addresses a primary research field: the automatic analysis of surgical workflows through the understanding of surgical activities from videos. While a large body of research has focused on recognizing activities at a single specific level of granularity—such as phases, steps, or actions—the principal theme of this thesis is the multi-level understanding of surgical activities.

First, I addressed the foundational need for high-quality data by constructing a comprehensive, multi-level cataract surgery dataset. This dataset is meticulously annotated with synchronized labels for activities at three levels of granularity—phases, steps, and fine-grained actions—providing a robust benchmark for developing and validating the models proposed in this thesis.

Subsequently, I tackled the challenge of fine-grained surgical action detection. Recognizing that existing methods often struggle with accurate action classification and fail to provide model confidence, I proposed the Anchor-context Action Detection Network (ACTNet) framework. This network uses an anchor-context module to determine where and what actions occur and a conditional diffusion module to refine predictions and, crucially, to provide

a confidence score for each prediction. A limitation of this method is that its performance is dependent on the clear visibility of instruments, which can be compromised by visual challenges like surgical smoke, blur, or occlusion.

The following research addressed the need for a holistic, multi-level understanding of the surgical scene. To move beyond single-task models, I proposed a Hierarchical Context Transformer (HCT) framework that conceptualizes surgical activity as an interconnected hierarchy. The framework was designed to concurrently model the interactions between phases, steps, and actions to augment task-specific features. A minor limitation of this approach is its assumption of a relatively consistent hierarchical structure, which may not fully hold in atypical or highly complex surgical cases where the workflow deviates significantly from the standard procedure.

To enhance model robustness, I then tackled the critical but often-overlooked issue of spurious correlations, where models learn from dataset biases rather than true causal relationships. I introduced a causality-inspired framework for surgical video representation learning that uses causal intervention via backdoor adjustment to disentangle the model from confounding factors. This ensures the model learns true causal effects, leading to improved generalization. A challenge in this approach is that defining the correct causal graph is a non-trivial task that relies on expert knowledge; an incorrectly specified graph could potentially mislead the model.

Finally, this thesis took a step forward to address the critical challenge of generalization across different clinical environments. To this end, I developed a comprehensive foundation model for multi-level workflow recognition in cataract surgery, specifically designed for robust performance across multiple medical centers. Extensive multi-institutional evaluations demonstrated the model's ability to reliably adapt to inter-center variability. As identified in our study, one limitation is that while temporal features enhance performance on intra-dataset tasks, they can sometimes encode dataset-specific biases that may hinder cross-center transferability, suggesting the need for context-dependent strategies.

The problems addressed in this thesis have many potential applications both in and out of the OR, such as objective skill assessment, automated surgical report generation, real-time safety monitoring, and eventually instigating autonomous surgical systems. Beyond these applications, this thesis provides several specific and direct contributions to the wider research community: (1) a new, comprehensive, multi-level cataract surgery dataset that enables the development and benchmarking of future hierarchical models; (2) a suite of novel, validated frameworks (ACTNet, HCT, DSVRL, and the COS Model) that serve as strong baselines for fine-grained detection, hierarchical modeling, causal deconfounding, and generalization; and (3) a set of key insights demonstrating that explicitly modeling context, causality, and domain generalization is critical for building robust surgical AI systems. These contributions are intended to provide foundational tools and considerably influence future research in surgical data science, specifically on multi-level activity recognition.

8.2 Future Work

This thesis has laid a solid foundation for multi-level surgical scene understanding using video data. However, there are several promising directions for future research that could further advance the field.

Weakly-supervised and Self-supervised Learning. A primary challenge highlighted in this thesis is the creation of large-scale, meticulously annotated datasets. The frame-by-frame labeling process is both time-consuming and requires extensive expert knowledge. To address this bottleneck, future work could explore weakly-supervised learning approaches. For instance, models could be trained to recognize fine-grained steps and actions using only high-level phase labels as a weaker form of supervision. Furthermore, self-supervised learning presents a powerful alternative. By leveraging vast amounts of unlabeled surgical videos, it is possible to pre-train robust models that learn meaningful spatial and temporal representations of surgical procedures. These learned representations could then be fine-tuned for

various downstream tasks with a much smaller amount of labeled data.

Few-shot Learning and Domain Adaptation. While this thesis successfully developed a foundation model demonstrating strong generalization across different medical centers, adapting such models to new, unseen environments (e.g., a new hospital, surgeon, or equipment) with minimal data remains a challenge. Creating large labeled datasets for every possible variation is impractical. Consequently, few-shot learning emerges as a crucial research direction. Future studies could focus on developing methods that allow the foundation model to quickly adapt to a new domain using only a handful of labeled video examples. This would greatly enhance the model's practical applicability and scalability in real-world clinical settings.

Integration with Large Language Models. The current work focuses exclusively on visual data. A significant opportunity lies in creating multimodal systems that combine video understanding with the power of Large Language Models (LLMs). Such a framework could enable novel applications, for example, automatically generating detailed surgical reports from video feeds in real-time. It could also support an interactive query system where surgeons could ask questions in natural language (e.g., "How long did the phacoemulsification step take?") and receive instant answers. Moreover, textual surgical manuals and protocols could be used as an additional source of knowledge to guide and regularize the video analysis models.

Multimodal Frameworks with Non-visual Data. Surgical procedures generate a wealth of data beyond the video stream. As mentioned in the introduction of this thesis, data from robotic systems (kinematics), patient physiological monitors (e.g., heart rate, blood pressure), and other sensors are also available. Future research should aim to develop a comprehensive multimodal framework that integrates these different data streams. By fusing video information with instrument movements and patient vital signs, the system could achieve a more holistic understanding of the surgical context. This could lead to more robust safety systems capable of predicting not only procedural events but also potential

adverse patient reactions.

Interactive Surgical Training and Assessment Systems. A key application of this research is surgical training. The next logical step is to embed the developed AI models into a complete surgical training system. This involves integrating the multi-level recognition algorithms with hardware, such as physical simulators or Augmented Reality (AR) headsets. Such a system could provide trainees with real-time, objective feedback on their performance, highlighting deviations from standard procedures or inefficient movements. By offering a data-driven approach to skill assessment, this would move beyond subjective evaluation and contribute to a more standardized and effective surgical education.

Appendix One

Supplementary for Chapter 3

Table A.1: Definitions of all the proposed phases for the cataract surgery procedure.

Phase ID	Phase Name	Description
P0	No phase	No defined surgical phase is currently active or identified.
P1	Opening	The initial surgical stage focused on meticulously preparing the eye and establishing sterile access to the anterior segment for subsequent maneuvers.
P2	Phacoemulsification	The core stage involving the precise ultrasonic fragmentation (phacoemulsification) of the opaque lens nucleus and cortex, followed by its complete aspiration from the capsular bag.
P3	Implantation	The critical stage of accurately inserting and positioning the artificial intraocular lens (IOL) within the empty capsular bag or, if necessary, in an alternative stable location.
P4	Closure	The final surgical stage dedicated to ensuring watertight wound closure, restoring normal intraocular pressure, and completing all necessary post-operative checks before concluding the procedure.

A.1 Label Definition for Cataract Dataset

The definitions for phases, steps and actions are presented in Table A.1 (Phases), Table A.2 (Steps), and Table A.3 (Actions), respectively.

The phase level offers a high-level summary of the surgery's main stages, such as Open-

ing, Phacoemulsification, Implantation, and Closure. Among these, the Phacoemulsification phase, involving ultrasonic fragmentation and aspiration of the cataractous lens, is the most crucial. It enables minimally invasive surgery through micro-incisions, leading to reduced trauma, faster visual recovery, and has become the gold standard for cataract extraction. The step level further refines these phases into more detailed surgical procedures. Critical steps for successful outcomes include precise Incision construction, the creation of a continuous curvilinear Capsulorhexis, the Phacoemulsification step itself (encompassing nucleus management), and accurate Lens Implantation. At the most granular level, actions describe moment-to-moment operations, detailing the instrument used and its interaction with ocular tissues. For instance, an action might specify the use of capsulorhexis forceps to tear the anterior capsule.

Table A.2: Definitions of all the proposed steps for the cataract surgery procedure.

Step ID	Step Name	Description
S0	No step	No specific surgical step is currently being performed or identified within a phase.
S1	Incision	The precise creation of one or more surgical openings through the cornea or sclera to allow entry of instruments and fluids into the anterior chamber.
S2	Viscous-agent-injection	The controlled injection of a viscoelastic substance into the anterior chamber to maintain its depth, protect delicate intraocular structures like the corneal endothelium, and facilitate surgical maneuvers.

S3	Rhexis	The meticulous creation of a continuous, curvilinear opening (capsulorhexis) in the anterior lens capsule, providing access to the lens nucleus and cortex for removal.
S4	Hydrodissection	The gentle injection of balanced salt solution beneath the anterior lens capsule to carefully separate the lens nucleus and epinucleus from the surrounding cortex and posterior capsule, facilitating easier rotation and removal.
S5	Phacoemulsification	The primary cataract removal step involving the use of ultrasonic energy to break down (emulsify) the hardened lens nucleus into smaller fragments, which are then simultaneously aspirated from the eye.
S6	Irrigation-and-aspiration	The thorough removal of any remaining soft lens cortex material from the capsular bag using a combination of controlled irrigation to dislodge particles and aspiration to remove them.
S7	Capsule-polishing	The careful cleaning and polishing of the inner surfaces of the lens capsule, particularly the posterior capsule, to remove residual lens epithelial cells and debris, reducing the risk of posterior capsule opacification.
S8	Lens-implant	The precise insertion of the selected artificial IOL through the main incision and its careful unfolding and positioning within the capsular bag to restore vision.

S9	Viscous-agent-removal	The complete and careful aspiration of the previously injected viscoelastic substance from the anterior chamber and capsular bag to prevent post-operative intraocular pressure spikes.
S10	Wound-hydration	The application of balanced salt solution to the edges of corneal incisions to induce stromal hydration and swelling, promoting self-sealing of the wounds without sutures.

These phases, steps, and actions are hierarchically related, reflecting the surgical workflow. A single phase typically encompasses multiple steps, and each step may consist of several actions. For example, the Opening phase generally includes steps such as creating the incision, injecting a viscous agent, and performing the rhexis. These steps, involve specific actions like an incision knife cutting the corneal limbus, a cannula injecting viscoelastic material, or capsulorhexis forceps tearing the anterior capsule. While there is a general sequence, certain steps, like viscous agent injection, or actions, such as using a cotton swab to wipe the conjunctiva, may occur at various points throughout the procedure as deemed necessary by the surgeon.

Table A.3: Definitions of selected actions for the cataract surgery procedure.

Action ID	Action Name	Description
A0	No_action	No specific surgical action is being performed by an instrument or the surgeon at this moment; a state of pause or transition.

A1	Forceps_null_null	Forceps are present within the surgical field but are currently idle and not engaged in a specific manipulative task.
A2	Forceps_grasp_conjunctiva	Forceps are utilized to securely grasp the conjunctiva, often for stabilization or retraction during surgical exposure.
A3	Forceps_loosen_conjunctiva	Forceps are employed to gently dissect or separate the conjunctival tissue from underlying structures, aiding in exposure.
A4	Forceps_hold_needle	Forceps are used to firmly hold a suture needle, enabling precise passage through tissues for wound closure.
A5	Forceps_hold_wire	Forceps are engaged in gripping a wire, potentially for fixation, traction, or manipulation of an intraocular device.
A6	Forceps_pull_wire	Forceps are applying tensile force to a wire, typically to adjust its position or secure an implant.
A7	Forceps_slack_wire	Forceps are used to reduce tension on a wire, allowing for repositioning or removal.
A8	Incision-knife_null_null	An incision knife is present in the surgical field but is not actively being used to create an incision.
A9	Incision-knife_cut_corneal-limbus	An incision knife is precisely creating a surgical cut at the corneal-limbal junction to access the anterior chamber.

A10	Cannula_null_null	A cannula is positioned within the surgical field but is not currently being used for fluid injection or aspiration.
A11	Cannula_inject_saline	A cannula is delivering a stream of saline solution, typically for irrigation, hydration, or maintaining chamber stability.
A12	Cannula_inject_viscoelastic	A cannula is injecting a clear, gel-like viscoelastic substance to protect intraocular tissues and create surgical space.
A13	Cannula_inject_capsule-stain	A cannula is introducing a specialized dye to stain the anterior lens capsule, enhancing its visibility for capsulorhexis.
A14	Cannula_separate_lens-capsule	A cannula is directing a fluid jet (hydrodissection) to gently separate the lens capsule from the underlying lens material.
A15	Cannula_rotate_lens-nucleus	A cannula tip is being used to carefully manipulate or rotate the lens nucleus within the capsular bag.
A16	Cannula_seal_corneal-limbus	A cannula is applying fluid to the corneal-limbal incision site to hydrate the stroma and promote self-sealing.
A17	Capsulorhexis-forceps_null_null	Capsulorhexis forceps are available in the surgical field but are not currently engaged in manipulating the lens capsule.

A18	Capsulorhexis-forceps_tear_anterior-capsule	Capsulorhexis forceps are initiating or propagating a tear in the anterior lens capsule to create the capsulorhexis.
A19	Capsulorhexis-forceps_loosen_anterior-capsule	Capsulorhexis forceps are used to gently unfold, reposition, or free a flap of the anterior lens capsule during rhexis creation.
A20	Capsulorhexis-forceps_grasp_anterior-capsule	Capsulorhexis forceps are securely gripping the edge of the anterior lens capsule flap to guide the capsulorhexis.
A21	Phacoemulsifier_null_null	The phacoemulsification handpiece is present and powered on but is not currently delivering ultrasonic energy or aspirating.
A22	Phacoemulsifier_inject_saline	The phacoemulsification handpiece is irrigating the anterior chamber with saline solution to maintain clarity and pressure.
A23	Phacoemulsifier_rotate_lens-nucleus	The tip of the phacoemulsification handpiece is being used as an instrument to rotate the lens nucleus for better access.
A24	Phacoemulsifier_push_lens-nucleus	The tip of the phacoemulsification handpiece is gently pushing or nudging the lens nucleus for repositioning within the capsular bag.
A25	Phacoemulsifier_chop_lens-nucleus	The phacoemulsification handpiece is actively delivering ultrasonic energy to break down (chop or sculpt) the lens nucleus into manageable fragments.

A26	Phacoemulsifier_aspirate_lens-nucleus	The phacoemulsification handpiece is aspirating the emulsified lens nucleus fragments from the eye while maintaining irrigation.
A27	Lens-hook_null_null	A lens hook or manipulator is present in the surgical field but is not currently being used to engage ocular structures.
A28	Lens-hook_push_lens-nucleus	A lens hook is applying gentle pressure to push or reposition the lens nucleus within the capsular bag.
A29	Lens-hook_pull_lens-nucleus	A lens hook is used to gently pull or retract the lens nucleus, often to aid in rotation or fragmentation.
A30	Lens-hook_grasp_lens-nucleus	A lens hook is engaging a part of the lens nucleus to provide a point of leverage or stabilization.
A31	Lens-hook_push_intraocular-lens	A lens hook is used to delicately nudge or push the intraocular lens to achieve optimal centration or orientation.
A32	Lens-hook_pull_intraocular-lens	A lens hook is used to gently pull or adjust the position of the intraocular lens or its haptics within the eye.
A33	Cotton-swab_null_null	A sterile cotton swab is available but is not currently being applied to any ocular or adnexal surface.

A34	Cotton-swab_wipe_conjunctiva	A cotton swab is used to gently wipe or absorb fluid/debris from the conjunctival surface or fornices.
A35	Cotton-swab_wipe_eyelid	A cotton swab is used to clean or dry the eyelid margins or surrounding skin area.
A36	Irrigation-aspiration-handpiece_null_null	The irrigation-aspiration (I/A) handpiece is present and connected but is not actively irrigating or aspirating.
A37	Irrigation-aspiration-handpiece_inject_saline	The I/A handpiece is delivering a controlled flow of saline solution into the eye for irrigation purposes.
A38	Irrigation-aspiration-handpiece_aspirate_lens-cortex	The I/A handpiece is actively aspirating residual lens cortex material from the capsular bag while simultaneously irrigating.
A39	Irrigation-aspiration-handpiece_polish_intraocular-lens	The I/A handpiece, often with a specialized tip, is used to polish the posterior capsule or the surface of the intraocular lens.
A40	Aspiration-handpiece_null_null	An aspiration-only handpiece is present but is not currently applying suction.
A41	Aspiration-handpiece_aspirate_lens-cortex	An aspiration-only handpiece is applying suction to remove lens cortex material from the capsular bag.
A42	Aspiration-handpiece_polish_intraocular-lens	An aspiration-only handpiece, possibly with a polishing tip, is used to clean the surface of the intraocular lens or posterior capsule.

A43	Irrigation-handpiece_null_null	An irrigation-only handpiece is present but is not currently delivering fluid.
A44	Irrigation-handpiece_inject_saline	An irrigation-only handpiece is delivering a flow of saline solution into the eye, typically for chamber maintenance or gentle dissection.
A45	Implant-injector_null_null	An intraocular lens implant injector system is prepared but the lens has not yet been injected.
A46	Implant-injector_inject_intraocular-lens	The implant injector system is being used to controllably deliver and unfold the intraocular lens into the eye.
A47	Scissors_null_null	Surgical scissors are present in the field but are not currently being used to cut any material.
A48	Scissors_snip_wire	Surgical scissors are precisely cutting a wire, such as a suture or a component of an intraocular device.
A49	Scissors_snip_fibre	Surgical scissors are used to snip a fibre, most commonly a suture thread, or occasionally other fine materials.

References

[1] Paul R Barach and Bill Rostenberg. "Design of cardiac surgery operating rooms and the impact of the built environment". In: *Pediatric and Congenital Cardiac Care: Volume 2: Quality Improvement and Patient Safety*. Springer, 2014, pp. 411–424.

[2] Shubham Bobade and Sheetal Asutkar. "Current trends and future directions in surgery: A brief scoping review". In: *Multidisciplinary Reviews* 8.1 (2025), pp. 2025028–2025028.

[3] Apollon Zygomalas, Dimitris Kalles, and George Skroubis. "The role of Artificial Intelligence in modern surgery". In: *Achaiki Iatriki* (2020), p. 63.

[4] Andrea Patricia Pérez Ayme, Josseline María Caiza Suárez, Martin Mateo Paredes Ortega, Gualoto Gualoto Gualoto Gualoto, Santillán Lima Santillán Lima, Adriana Elizabeth Rivera Campoverde, Alfredo Augusto Rivera Ticona, Coello Vergara Coello Vergara, and Gerzon David Martínez Serrano. "Advancements in minimally invasive surgical techniques: A comprehensive review". In: *Salud, Ciencia y Tecnología* 4 (2024), pp. 745–745.

[5] Raphael Uwechue, Pankaj Chandak, Zubir Ahmed, Petrut Gogalniceanu, Nicos Kessaris, and Nizam Mamode. "Minimally invasive surgical techniques for kidney transplantation". In: *The Cochrane Database of Systematic Reviews* 2017.7 (2017), p. CD012698.

[6] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. "Surgical data science for next-generation interventions". In: *Nature Biomedical Engineering* 1.9 (2017), pp. 691–696.

[7] Kun Yuan, Nassir Navab, Nicolas Padoy, et al. "Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 122952–122983.

[8] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. "Surgical data science–from concepts toward clinical translation". In: *Medical Image Analysis* 76 (2022), p. 102306.

[9] Lena Maier-Hein, Matthias Eisenmann, Carolin Feldmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, Bernard Gibaud, Gregory D Hager, Makoto Hashizume, Darko Katic, et al. "Surgical data science: A consensus perspective". In: *arXiv preprint arXiv:1806.03184* (2018).

[10] Romina Pedrett, Pietro Mascagni, Guido Beldi, Nicolas Padoy, and Joël L Lavanchy. "Technical skill assessment in minimally invasive surgery using artificial intelligence: A systematic review". In: *Surgical Endoscopy* 37.10 (2023), pp. 7412–7424.

[11] Beatrice van Amsterdam, Matthew J Clarkson, and Danail Stoyanov. "Gesture recognition in robotic surgery: a review". In: *IEEE Transactions on Biomedical Engineering* 68.6 (2021).

[12] Tom Vercauteren, Mathias Unberath, Nicolas Padoy, and Nassir Navab. "Cai4cai: The rise of contextual artificial intelligence in computer-assisted interventions". In: *Proceedings of the IEEE* 108.1 (2019), pp. 198–214.

[13] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. "Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 593–603.

[14] Fangqiu Yi and Tingting Jiang. "Hard frame detection and online mapping for surgical phase recognition". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 449–457.

[15] Mingxiao Tu, Hoijoon Jung, Jinman Kim, and Andre Kyme. "Head-mounted displays in context-aware systems for open surgery: A state-of-the-art review". In: *IEEE Journal of Biomedical and Health Informatics* (2024).

[16] Yueming Jin, Yang Yu, Cheng Chen, Zixu Zhao, Pheng-Ann Heng, and Danail Stoyanov. "Exploring intra-and inter-video relation for surgical semantic scene segmentation". In: *IEEE Transactions on Medical Imaging* 41.11 (2022), pp. 2991–3002.

[17] Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. "Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos". In: *Scientific Data* 11.1 (2024), p. 373.

[18] Shu Yang, Zhiyuan Cai, Luyang Luo, Ning Ma, Shuchang Xu, and Hao Chen. "Surg-petl: Parameter-efficient image-to-surgical-video transfer learning for surgical phase recognition". In: *arXiv preprint arXiv:2409.20083* (2024).

[19] Tongbiao Cai and Zijian Zhao. "Convolutional neural network-based surgical instrument detection". In: *Technology and Health Care* 28.1_suppl (2020), pp. 81–88.

[20] Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio

Caicedo, Nicolás Fernández, et al. "Pixel-wise recognition for holistic surgical scene understanding". In: *arXiv preprint arXiv:2401.11174* (2024).

[21] Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. "Towards holistic surgical scene understanding". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 442–452.

[22] Kubilay Can Demir, Hannah Schieber, Tobias Weise, Daniel Roth, Matthias May, Andreas Maier, and Seung Hee Yang. "Deep learning in surgical workflow analysis: A review of phase and step recognition". In: *IEEE Journal of Biomedical and Health Informatics* 27.11 (2023), pp. 5405–5417.

[23] Danyal Z Khan, Imanol Luengo, Santiago Barbarisi, Carole Addis, Lucy Culshaw, Neil L Dorward, Pinja Haikka, Abhiney Jain, Karen Kerr, Chan Hee Koh, et al. "Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: Development and preclinical evaluation (IDEAL stage 0)". In: *Journal of Neurosurgery* 137.1 (2021), pp. 51–58.

[24] Heba Taher, Vincent Grasso, Sherifa Tawfik, and Andrew Gumbs. "The challenges of deep learning in artificial intelligence and autonomous actions in surgery: A literature review". In: *Artificial Intelligence Surgery* 2.3 (2022), pp. 144–158.

[25] Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy. "Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures". In: *International Journal of Computer Assisted Radiology and Surgery* 16 (2021), pp. 1111–1119.

[26] Ozanan R Meireles, Guy Rosman, Maria S Altieri, Lawrence Carin, Gregory Hager, Amin Madani, Nicolas Padoy, Carla M Pugh, Patricia Sylla, Thomas M Ward, et al.

"SAGES consensus recommendations on an annotation framework for surgical video". In: *Surgical Endoscopy* 35.9 (2021), pp. 4918–4929.

[27] Maryam Gholinejad, Arjo J. Loeve, and Jenny Dankelman and. "Surgical process modelling strategies: Which method to choose for determining workflow?" In: *Minimally Invasive Therapy & Allied Technologies* 28.2 (2019), pp. 91–104.

[28] Oliver Weede, Frank Dittrich, Heinz Wörn, Brian Jensen, Alois Knoll, Dirk Wilhelm, Michael Kranzfelder, Armin Schneider, and Hubertus Feussner. "Workflow analysis and surgical phase recognition in minimally invasive surgery". In: *2012 IEEE International Conference on Robotics and Biomimetics*. IEEE. 2012, pp. 1080–1074.

[29] Carlos Federico Davrieux, Mariano Palermo, Edgardo Serra, Eduardo Javier Houghton, Pablo Agustín Acquafresca, Caetano Finger, and Mariano Eduardo Giménez. "Stages and factors of the "perioperative process": Points in common with the aeronautical industry". In: *ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)* 32 (2019), e1423.

[30] Katia Charrière, Gwénolé Quellec, Mathieu Lamard, David Martiano, Guy Cazuguel, Gouenou Coatrieux, and Béatrice Cochener. "Real-time analysis of cataract surgery videos using statistical models". In: *Multimedia Tools and Applications* 76 (2017), pp. 22473–22491.

[31] Antonio Baldascino, Matteo Mario Carlà, Federico Giannuzzi, Francesco Boselli, Tomaso Caporossi, Gloria Gambini, Antonio Villano, Aldo Caporossi, and Stanislao Rizzo. "Femtosecond laser-assisted cataract surgery: Analysis of surgical phases and comparison with standard phacoemulsification in uncomplicated cataracts". In: *Vision* 6.4 (2022), p. 72.

[32] Simon Müller, Mohit Jain, Bhuvan Sachdeva, Payal N Shah, Frank G Holz, Robert P Finger, Kaushik Murali, Maximilian WM Wintergerst, and Thomas Schultz. "Arti-

ficial intelligence in cataract surgery: A systematic review". In: *Translational Vision Science & Technology* 13.4 (2024), pp. 20–20.

[33] Kadir Kirtac, Nizamettin Aydin, Joël L Lavanchy, Guido Beldi, Marco Smit, Michael S Woods, and Florian Aspart. "Surgical phase recognition: From public datasets to real-world data". In: *Applied Sciences* 12.17 (2022), p. 8746.

[34] Carmen Guzmán-García, Patricia Sánchez-González, Ignacio Oropesa, and Enrique J Gómez. "Automatic assessment of procedural skills based on the surgical workflow analysis derived from speech and video". In: *Bioengineering* 9.12 (2022), p. 753.

[35] Fabio Carrillo, Hooman Esfandiari, Sandro Müller, Marco von Atzigen, Aidana Massalimova, Daniel Suter, Christoph J Laux, José M Spirig, Mazda Farshad, and Philipp Fürnstahl. "Surgical process modeling for open spinal surgeries". In: *Frontiers in Surgery* 8 (2022), p. 776945.

[36] Georgii Kostiuchik, Lalith Sharan, Benedikt Mayer, Ivo Wolf, Bernhard Preim, and Sandy Engelhardt. "Surgical phase and instrument recognition: How to identify appropriate dataset splits". In: *International Journal of Computer Assisted Radiology and Surgery* 19.4 (2024), pp. 699–711.

[37] Zixin Chen, Dewei Yang, Ang Li, Louzong Sun, Jifan Zhao, Jie Liu, Linxun Liu, Xiaobo Zhou, Yonghua Chen, Yunqiang Cai, et al. "Decoding surgical skill: an objective and efficient algorithm for surgical skill classification based on surgical gesture features–experimental studies". In: *International Journal of Surgery* 110.3 (2024), pp. 1441–1449.

[38] Nicolas Padoy. "Workflow and activity modeling for monitoring surgical procedures". PhD thesis. Université Henri Poincaré-Nancy 1; Technische Universität München, 2010.

[39] Zhuohong He, Ali Mottaghi, Aidean Sharghi, Muhammad Abdullah Jamal, and Omid Mohareri. "An empirical study on activity recognition in long surgical videos". In: *Machine Learning for Health*. PMLR. 2022, pp. 356–372.

[40] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. "Endonet: a deep architecture for recognition tasks on laparoscopic videos". In: *IEEE Transactions on Medical Imaging* 36.1 (2016), pp. 86–97.

[41] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maršalkaitė, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, et al. "Cataracts: Challenge on automatic tool annotation for cataract surgery". In: *Medical Image Analysis* 52 (2019), pp. 24–41.

[42] Odysseas Zisimopoulos, Evangello Flouty, Imanol Luengo, Petros Giataganas, Jean Nehme, Andre Chow, and Danail Stoyanov. "Deepphase: surgical phase recognition in cataracts videos". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 265–272.

[43] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, et al. "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark". In: *Medical Image Analysis* 86 (2023), p. 102770.

[44] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos". In: *Medical Image Analysis* 78 (2022), p. 102433.

[45] Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, et

al. “Micro-surgical anastomose workflow recognition challenge report”. In: *Computer Methods and Programs in Biomedicine* 212 (2021), p. 106452.

[46] Arthur Derathé, Fabian Reche, Sylvain Guy, Katia Charrière, Bertrand Trilling, Pierre Jannin, Alexandre Moreau-Gaudry, Bernard Gibaud, and Sandrine Voros. “Lapex: A new multimodal dataset for context recognition and practice assessment in laparoscopic surgery”. In: *Scientific Data* 12.1 (2025), p. 342.

[47] Joël L Lavanchy, Sanat Ramesh, Diego Dall’Alba, Cristians Gonzalez, Paolo Fiorini, Beat P Müller-Stich, Philipp C Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. “Challenges in multi-centric generalization: Phase and step recognition in Roux-en-Y gastric bypass surgery”. In: *International Journal of Computer Assisted Radiology and Surgery* (2024), pp. 1–9.

[48] Maria Robu, Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. “Towards real-time multiple surgical tool tracking”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 9.3 (2021), pp. 279–285.

[49] Ufaq Khan, Umair Nawaz, Adnan Qayyum, Shazad Ashraf, Muhammad Bilal, and Junaid Qadir. “Surgical scene understanding in the era of foundation AI models: A comprehensive review”. In: *arXiv preprint arXiv:2502.14886* (2025).

[50] Haozheng Xu and Stamatia Giannarou. “Occlusion-robust markerless surgical instrument pose estimation”. In: *Healthcare Technology Letters* 11.6 (2024), pp. 327–335.

[51] Ranyang Li, Junjun Pan, Yaqing Si, Bin Yan, Yong Hu, and Hong Qin. “Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition”. In: *IEEE Transactions on Medical Imaging* 39.2 (2019), pp. 328–340.

[52] Shintaro Arakaki, Shin Takenaka, Kimimasa Sasaki, Daichi Kitaguchi, Hiro Hasegawa, Nobuyoshi Takeshita, Mitsuhisa Takatsuki, and Masaaki Ito. “Artificial intelligence

in minimally invasive surgery: Current state and future challenges". In: *JMA Journal* 8.1 (2025), pp. 86–90.

[53] Qingyuan Zheng, Rui Yang, Xinmiao Ni, Song Yang, Zhengyu Jiang, Lei Wang, Zhiyuan Chen, and Xiuheng Liu. "Development and validation of a deep learning-based laparoscopic system for improving video quality". In: *International Journal of Computer Assisted Radiology and Surgery* 18.2 (2023), pp. 257–268.

[54] Yakui Chu, Xilin Yang, Heng Li, Danni Ai, Yuan Ding, Jingfan Fan, Hong Song, and Jian Yang. "Multi-level feature aggregation network for instrument identification of endoscopic images". In: *Physics in Medicine & Biology* 65.16 (2020), p. 165004.

[55] Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. "Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning". In: *arXiv preprint arXiv:2408.07931* (2024).

[56] Zijian Zhao, Zhaorui Chen, Sandrine Voros, and Xiaolin Cheng. "Real-time tracking of surgical instruments based on spatio-temporal context and deep learning". In: *Computer Assisted Surgery* 24.sup1 (2019), pp. 20–29.

[57] Courtney A Green, Patricia S O'Sullivan, Ankit Sarin, and Hueylan Chern. "Micro-analysis of video from a robotic surgical procedure: implications for observational learning in the robotic environment". In: *Journal of Robotic Surgery* 13 (2019), pp. 449–454.

[58] Nicola Rieke, David Joseph Tan, Chiara Amat di San Filippo, Federico Tombari, Mohamed Alsheakhali, Vasileios Belagiannis, Abouzar Eslami, and Nassir Navab. "Real-time localization of articulated surgical instruments in retinal microsurgery". In: *Medical Image Analysis* 34 (2016), pp. 82–100.

[59] Veronica Penza, Sara Moccia, Elena De Momi, and Leonardo S Mattos. "Enhanced vision to improve safety in robotic surgery". In: *Handbook of Robotic and Image-Guided Surgery*. Elsevier, 2020, pp. 223–237.

[60] Ling Li, Xiaojian Li, Shuai Ding, Zhao Fang, Mengya Xu, Hongliang Ren, and Shanlin Yang. "Sirnet: Fine-grained surgical interaction recognition". In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 4212–4219.

[61] Yunlong Li, Zijian Zhao, Renbo Li, and Feng Li. "Deep learning for surgical workflow analysis: A survey of progresses, limitations, and trends". In: *Artificial Intelligence Review* 57.11 (2024), p. 291.

[62] Haiqi Wei, Frank Rudzicz, David Fleet, Teodor Grantcharov, and Babak Taati. "Intraoperative adverse event detection in laparoscopic surgery: Stabilized multi-stage temporal convolutional network with focal-uncertainty loss". In: *Machine Learning for Healthcare Conference*. PMLR. 2021, pp. 283–307.

[63] Xinpeng Ding, Xinjian Yan, Zixun Wang, Wei Zhao, Jian Zhuang, Xiaowei Xu, and Xiaomeng Li. "Less is more: Surgical phase recognition from timestamp supervision". In: *IEEE Transactions on Medical Imaging* 42.6 (2023), pp. 1897–1910.

[64] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 343–352.

[65] Yanzhe Liu, Shang Zhao, Gong Zhang, Xiuping Zhang, Minggen Hu, Xuan Zhang, Chenggang Li, S Kevin Zhou, and Rong Liu. "Multilevel effective surgical workflow recognition in robotic left lateral sectionectomy with deep learning: experimental research". In: *International Journal of Surgery* 109.10 (2023), pp. 2941–2952.

[66] Yanqi Cheng, Lihao Liu, Shujun Wang, Yueming Jin, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. "Why deep surgical models fail?: Revisiting surgical action triplet recognition through the lens of robustness". In: *International Workshop on Trustworthy Machine Learning for Healthcare*. Springer. 2023, pp. 177–189.

[67] Yuanhao Zhai, Ziyi Liu, Zhenyu Wu, Yi Wu, Chunluan Zhou, David Doermann, Junsong Yuan, and Gang Hua. "Soar: Scene-debiasing open-set action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 10244–10254.

[68] Michael Gaies, Mary K Olive, Gabe E Owens, John R Charpie, Wenying Zhang, Sara K Pasquali, Darren Klugman, John M Costello, Steven M Schwartz, and Mousumi Banerjee. "Methods to enhance causal inference for assessing impact of clinical informatics platform implementation". In: *Circulation: Cardiovascular Quality and Outcomes* 16.2 (2023), e009277.

[69] Ali Mottaghi, Aidean Sharghi, Serena Yeung, and Omid Mohareri. "Adaptation of surgical activity recognition models across operating rooms". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 530–540.

[70] Lalithkumar Seenivasan, Mobarakol Islam, Chi-Fai Ng, Chwee Ming Lim, and Hongliang Ren. "Biomimetic incremental domain generalization with a graph network for surgical scene understanding". In: *Biomimetics* 7.2 (2022), p. 68.

[71] Amila Cizmic, Anuja T Mitra, Anas A Preukschas, Marius Kemper, Nathaniel T Melling, Oliver Mann, Sheraz Markar, Thilo Hackert, and Felix Nickel. "Artificial intelligence for intraoperative video analysis in robotic-assisted esophagectomy". In: *Surgical Endoscopy* (2025), pp. 1–10.

[72] Thomas F Byrd IV and Christopher J Tignanelli. "Artificial intelligence in surgery—a narrative review". In: *Journal of Medical Artificial Intelligence* 7 (2024).

[73] Chidozie N Ogbonnaya, Shizhou Li, Changshi Tang, Baobing Zhang, Paul Sullivan, Mustafa Suphi Erden, and Benjie Tang. "Exploring the role of artificial intelligence (AI)-driven training in laparoscopic suturing: A systematic review of skills mastery, retention, and clinical performance in surgical education". In: *Healthcare*. Vol. 13. 5. MDPI. 2025, p. 571.

[74] Kivanc Yangi, Thomas J On, Yuan Xu, Arianna S Gholami, Jinpyo Hong, Alexander G Reed, Pravarakhya Puppalla, Jiuxu Chen, Jonathan A Tangsrivimol, Baoxin Li, et al. "Artificial intelligence integration in surgery through hand and instrument tracking: A systematic literature review". In: *Frontiers in Surgery* 12 (2025), p. 1528362.

[75] Julian Varas, Brandon Valencia Coronel, IGNACIO VILLAGRáN, Gabriel Escalona, Rocio Hernandez, Gregory Schuit, VALENTINA DURáN, Antonia Lagos-Villaseca, Cristian Jarry, Andres Neyem, et al. "Innovations in surgical training: exploring the role of artificial intelligence and large language models (LLM)". In: *Revista do Colégio Brasileiro de Cirurgiões* 50 (2023), e20233605.

[76] Emmett D Goodman, Krishna K Patel, Yilun Zhang, William Locke, Chris J Kennedy, Rohan Mehrotra, Stephen Ren, Melody Guan, Orr Zohar, Maren Downing, et al. "Analyzing surgical technique in diverse open surgical videos with multitask machine learning". In: *JAMA Surgery* 159.2 (2024), pp. 185–192.

[77] Nir Messer, Eran Nizri, Guy Lahat, and Amir Szold. "Implementing an AI-powered endoscopic surgery video recording system in a large hospital network: lessons learned and future prospects". In: *Artificial Intelligence Surgery* 5.2 (2025), pp. 191–199.

[78] Lachlan Dick, Connor P Boyle, Richard JE Skipworth, Douglas S Smink, Victoria Ruth Tallentire, and Steven Yule. "Automated analysis of operative video in surgical training: Scoping review". In: *BJS Open* 8.5 (2024), zrae124.

[79] Niklas Pakkasjärvi, Tanvi Luthra, and Sachit Anand. "Artificial intelligence in surgical learning". In: *Surgeries* 4.1 (2023), pp. 86–97.

[80] Nasseh Hashemi, Matias Mose, Lasse R Østergaard, Flemming Bjerrum, Mostaan Hashemi, Morten Svendsen, Mikkel L Friis, Martin G Tolsgaard, and Sten Rasmussen. "Video-based robotic surgical action recognition and skills assessment on porcine models using deep learning". In: *Surgical Endoscopy* (2025), pp. 1–11.

[81] Monica Ortenzi, Judith Rapoport Ferman, Alenka Antolin, Omri Bar, Maya Zohar, Ori Perry, Dotan Asselmann, and Tamir Wolf. "A novel high accuracy model for automatic surgical workflow recognition using artificial intelligence in laparoscopic totally extraperitoneal inguinal hernia repair (TEP)". In: *Surgical Endoscopy* 37.11 (2023), pp. 8818–8828.

[82] Alejandra Perez, Han Zhang, Yu-Chun Ku, Lalithkumar Seenivasan, Roger Soberanis, Jose L Porras, Richard Day, Jeff Jopling, Peter Najjar, and Mathias Unberath. "Privacy-preserving operating room workflow analysis using digital twins". In: *arXiv preprint arXiv:2504.12552* (2025).

[83] Daniel J Lee, James Ding, and Thomas J Guzzo. "Improving operating room efficiency". In: *Current Urology Reports* 20 (2019), pp. 1–8.

[84] Jennifer A Eckhoff, Guy Rosman, Maria S Altieri, Stefanie Speidel, Danail Stoyanov, Mehran Anvari, Lena Meier-Hein, Keno März, Pierre Jannin, Carla Pugh, et al. "SAGES consensus recommendations on surgical video data use, structure, and exploration (for research in artificial intelligence, clinical quality improvement, and surgical education)". In: *Surgical Endoscopy* 37.11 (2023), pp. 8690–8707.

[85] Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. "EndoChat: Grounded multimodal large language model for endoscopic surgery". In: *arXiv preprint arXiv:2501.11347* (2025).

[86] Yang Liu, Yu-Shen Wei, Hong Yan, Guan-Bin Li, and Liang Lin. "Causal reasoning meets visual representation learning: A prospective study". In: *Machine Intelligence Research* 19.6 (2022), pp. 485–511.

[87] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. "Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 486–496.

[88] Tobias Blum, Nicolas Padoy, Hubertus Feußner, and Nassir Navab. "Modeling and online recognition of surgical phases using hidden markov models". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2008, pp. 627–635.

[89] Chang Shi, Yi Zheng, and Ann Majewicz Fey. "Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2022, pp. 8017–8024.

[90] Long Bai, Guankun Wang, Jie Wang, Xiaoxiao Yang, Huxin Gao, Xin Liang, An Wang, Mobarakol Islam, and Hongliang Ren. "Ossar: Towards open-set surgical activity recognition in robot-assisted surgery". In: *2024 IEEE International Conference on Robotics and Automation*. IEEE. 2024, pp. 14622–14629.

[91] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. "Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network". In: *IEEE Transactions on Medical Imaging* 37.5 (2017), pp. 1114–1126.

[92] Tobias Czempiel, Magdalini Paschali, Daniel Ostler, Seong Tae Kim, Benjamin Busam, and Nassir Navab. "Opera: Attention-regularized transformers for surgical phase recognition". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 604–614.

[93] Alejandra Pérez, Santiago Rodríguez, Nicolás Ayobi, Nicolás Aparicio, Eugénie Dessevres, and Pablo Arbeláez. "Must: Multi-scale transformers for surgical phase recognition". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 422–432.

[94] Minyoung Park, Seungtaek Oh, Taikyeong Jeong, and Sungwook Yu. "Multi-stage temporal convolutional network with moment loss and positional encoding for surgical phase recognition". In: *Diagnostics* 13.1 (2022), p. 107.

[95] Xinpeng Ding and Xiaomeng Li. "Exploring segment-level semantics for online phase recognition from surgical videos". In: *IEEE Transactions on Medical Imaging* 41.11 (2022), pp. 3309–3319.

[96] Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. "Skit: A fast key information video transformer for online surgical phase recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Oct. 2023, pp. 21074–21084.

[97] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, et al. "The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods". In: *arXiv preprint arXiv:2104.03178* (2021).

[98] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in Neural Information Processing Systems* 28 (2015).

[99] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.

[100] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. "Recognition of instrument-tissue interactions in endoscopic videos via action triplets". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 364–374.

[101] Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. "Rendezvous in time: An attention-based temporal fusion approach for surgical triplet recognition". In: *International Journal of Computer Assisted Radiology and Surgery* 18.6 (2023), pp. 1053–1059.

[102] Wenjun Lin, Yan Hu, Luoying Hao, Dan Zhou, Mingming Yang, Huazhu Fu, Cheekong Chui, and Jiang Liu. "Instrument-tissue interaction quintuple detection in surgery videos". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 399–409.

[103] Wenjun Lin, Yan Hu, Huazhu Fu, Mingming Yang, Chin-Boon Chng, Ryo Kawasaki, Cheekong Chui, and Jiang Liu. "Instrument-tissue interaction detection framework for surgical video understanding". In: *IEEE Transactions on Medical Imaging* (2024).

[104] Oluwatosin Alabi, Tom Vercauteren, and Miaojing Shi. "Multitask learning in minimally invasive surgical vision: A review". In: *Medical Image Analysis* (2025), p. 103480.

[105] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. "Multi-task recurrent convolutional network with correlation loss for surgical video analysis". In: *Medical Image Analysis* 59 (2020), p. 101572.

[106] Lalithkumar Seenivasan, Sai Mitheran, Mobarakol Islam, and Hongliang Ren. "Global-reasoned multi-task learning model for surgical scene understanding". In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 3858–3865.

[107] Giorgia Marullo, Leonardo Tanzi, Luca Ulrich, Francesco Porpiglia, and Enrico Vezzetti. "A multi-task convolutional neural network for semantic segmentation and

event detection in laparoscopic surgery". In: *Journal of Personalized Medicine* 13.3 (2023), p. 413.

[108] Kexuan Zhang, Qiyu Sun, Chaoqiang Zhao, and Yang Tang. "Causal reasoning in typical computer vision tasks". In: *Science China Technological Sciences* 67.1 (2024), pp. 105–120.

[109] Judea Pearl. "Causal inference in statistics: An overview". In: *Statistics Surveys* 3 (2009), pp. 96–146.

[110] Judea Pearl. "An introduction to causal inference". In: *The International Journal of Biostatistics* 6.2 (2010).

[111] Judea Pearl. "The foundations of causal inference". In: *Sociological Methodology* 40.1 (2010), pp. 75–149.

[112] Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. "Causal inference meets deep learning: A comprehensive survey". In: *Research* 7 (2024), p. 0467.

[113] Dandan Li, Ziyu Guo, Qing Liu, Li Jin, Zequn Zhang, Kaiwen Wei, and Feng Li. "Click: Integrating causal inference and commonsense knowledge incorporation for counterfactual story generation". In: *Electronics* 12.19 (2023), p. 4173.

[114] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. "Long-tailed classification by keeping the good and removing the bad momentum causal effect". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1513–1524.

[115] Xiaoshuo Yan, Zhaochuan Li, Lei Meng, Zhuang Qi, Wei Wu, Zixuan Li, and Xiangxu Meng. "Empowering vision transformers with multi-scale causal intervention for long-tailed image classification". In: *arXiv preprint arXiv:2505.08173* (2025).

[116] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. "Causal intervention for weakly-supervised semantic segmentation". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 655–666.

[117] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. "Counterfactual samples synthesizing for robust visual question answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10800–10809.

[118] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. "Debiased visual question answering from feature and sample perspectives". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3784–3796.

[119] Jiangtong Li, Li Niu, and Liqing Zhang. "Knowledge proxy intervention for deconfounded video question answering". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2782–2793.

[120] Xu Yang, Hanwang Zhang, and Jianfei Cai. "Deconfounded image captioning: A causal retrospect". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.11 (2021), pp. 12996–13010.

[121] Xing Wu, Shaoqi Peng, Jingwen Li, Jian Zhang, Qun Sun, Weimin Li, Quan Qian, Yue Liu, and Yike Guo. "Causal inference in the medical domain: A survey". In: *Applied Intelligence* 54.6 (2024), pp. 4911–4934.

[122] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. "C-cam: Causal cam for weakly supervised semantic segmentation on medical image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11676–11685.

[123] Bowei Tian, Yexiao He, Meng Liu, Yucong Dai, Ziyao Wang, Shwai He, Guoheng Sun, Zheyu Shen, Wanghao Ye, Yongkai Wu, et al. "Fair diagnosis: Leveraging causal modeling to mitigate medical bias". In: *arXiv preprint arXiv:2412.04739* (2024).

[124] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. "Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning". In: *Frontiers in Artificial Intelligence* 5 (2022), p. 825565.

[125] Min Hun Lee and Chong Jun Chew. "Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023), pp. 1–22.

[126] Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Shen Zhao, Guanbin Li, Cheng-Lin Liu, and Liang Lin. "Cross-modal causal intervention for medical report generation". In: *arXiv preprint arXiv:2303.09117* (2023).

[127] Io Nam Wong, Olivia Monteiro, Daniel T Baptista-Hon, Kai Wang, Wenyang Lu, Zhuo Sun, Sheng Nie, and Yun Yin. "Leveraging foundation and large language models in medical artificial intelligence". In: *Chinese Medical Journal* 137.21 (2024), pp. 2529–2539.

[128] Ian A Scott and Guido Zuccon. "The new paradigm in machine learning–foundation models, large language models and beyond: A primer for physicians". In: *Internal Medicine Journal* 54.5 (2024), pp. 705–715.

[129] Yueyan Bian, Jin Li, Chuyang Ye, Xiuqin Jia, and Qi Yang. "Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models". In: *Chinese Medical Journal* 138.06 (2025), pp. 651–663.

[130] Ryan C Godwin, Avery Tung, Dan E Berkowitz, and Ryan L Melvin. "Transforming physiology and healthcare through foundation models". In: *Physiology* 40.3 (2025), pp. 233–246.

[131] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[132] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PmLR. 2021, pp. 8748–8763.

[133] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt". In: *International Journal of Machine Learning and Cybernetics* (2024), pp. 1–65.

[134] Kangxu Fan, Liang Liang, Hao Li, Weijun Situ, Wei Zhao, and Ge Li. "Research on medical image segmentation based on SAM and its future prospects". In: *Bioengineering* 12.6 (2025), p. 608.

[135] Yichi Zhang, Zhenrong Shen, and Rushi Jiao. "Segment anything model for medical image segmentation: Current applications and future directions". In: *Computers in Biology and Medicine* (2024), p. 108238.

[136] Sheng He, Rina Bao, Jingpeng Li, Patricia Ellen Grant, and Yangming Ou. "Accuracy of segment-anything model (SAM) in medical image segmentation tasks". In: *CoRR* (2023).

[137] Mudassar Ali, Tong Wu, Haoji Hu, Qiong Luo, Dong Xu, Weizeng Zheng, Neng Jin, Chen Yang, and Jincao Yao. "A review of the segment anything model (SAM) for medical image analysis: Accomplishments and perspectives". In: *Computerized Medical Imaging and Graphics* (2024), p. 102473.

[138] Zhiling Yan, Weixiang Sun, Rong Zhou, Zhengqing Yuan, Kai Zhang, Yiwei Li, Tianming Liu, Quanzheng Li, Xiang Li, Lifang He, et al. "Biomedical sam 2: Segment anything in biomedical images and videos". In: *arXiv preprint arXiv:2408.03286* (2024).

[139] Tingxuan Chen, Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. "Text-driven adaptation of foundation models for few-shot surgical workflow analysis". In: *arXiv preprint arXiv:2501.09555* (2025).

[140] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. "Recognition of instrument-tissue interactions in endoscopic videos via action triplets". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 364–374.

[141] Muhammad Ejaz Bashir, Nomaan Khan, and Maria Khalid. "Impact of training data size on model accuracy and computational efficiency in deep learning based medical image diagnosis". In: *Annual Methodological Archive Research Review* 3.4 (2025), pp. 313–326.

[142] Ying Yu, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang. "Clinical big data and deep learning: Applications, challenges, and future outlooks". In: *Big Data Mining and Analytics* 2.4 (2019), pp. 288–305.

[143] Xue-Wen Chen and Xiaotong Lin. "Big data deep learning: challenges and perspectives". In: *IEEE Access* 2 (2014), pp. 514–525.

[144] Mengfang Li, Yuanyuan Jiang, Yanzhou Zhang, and Haisheng Zhu. "Medical image analysis using deep learning algorithms". In: *Frontiers in Public Health* 11 (2023), p. 1273253.

[145] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[146] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2014, pp. 740–755.

[147] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[148] Alan Shiels and J Fielding Hejtmancik. "Biology of inherited cataracts and opportunities for treatment". In: *Annual Review of Vision Science* 5.1 (2019), pp. 123–149.

[149] Thomas M Ward, Pietro Mascagni, Amin Madani, Nicolas Padoy, Silvana Perretta, and Daniel A Hashimoto. "Surgical data science and artificial intelligence for surgical education". In: *Journal of Surgical Oncology* 124.2 (2021), pp. 221–230.

[150] Hye Yeon Yang, Seung Soo Hong, Jihun Yoon, Bokyung Park, Youngno Yoon, Dai Hoon Han, Gi Hong Choi, Min-Kook Choi, and Sung Hyun Kim. "Deep learning-based surgical phase recognition in laparoscopic cholecystectomy". In: *Annals of Hepato-Biliary-Pancreatic Surgery* 28.4 (2024), pp. 466–473.

[151] S Bodenstedt, S Speidel, M Wagner, J Chen, A Kisilenko, BP Müller, L Maier-Hein, B Oliveira, S Hong, J Zamora-Anaya, et al. "HeiChole surgical workflow analysis and

full scene segmentation (HeiSurF)". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021.

[152] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. "2017 robotic instrument segmentation challenge". In: *arXiv preprint arXiv:1902.06426* (2019).

[153] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. "2018 robotic scene segmentation challenge". In: *arXiv preprint arXiv:2001.11190* (2020).

[154] Abhimanyu S Ahuja, Alfredo A Paredes III, Mallory LS Eisel, Sejal Kodwani, Isabella V Wagner, Darby D Miller, and Syril Dorairaj. "Applications of artificial intelligence in cataract surgery: A review". In: *Clinical Ophthalmology* (2024), pp. 2969–2975.

[155] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenolé Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. "Cadis: Cataract dataset for surgical RGB-image segmentation". In: *Medical Image Analysis* 71 (2021), p. 102053.

[156] Zi Ye, Ru Zhou, Zili Deng, Dan Wang, Ying Zhu, Xiaoli Jin, Lijun Zhang, Tianxiang Chen, Hanwei Zhang, and Mingliang Wang. "A comprehensive video dataset for surgical laparoscopic action analysis". In: *Scientific Data* 12.1 (2025), pp. 1–10.

[157] Thomas M Ward, Danyal M Fer, Yutong Ban, Guy Rosman, Ozanan R Meireles, and Daniel A Hashimoto. "Challenges in surgical video annotation". In: *Computer Assisted Surgery* 26.1 (2021), pp. 58–68.

[158] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. "Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition". In: *Interna-*

tional Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2024, pp. 306–316.

[159] Rachid Tahiri, Cati Albou-Ganem, Damien Gatinel, and Otman Sandali. "Innovations in cataract surgery". In: *Frontiers in Medicine* 11 (2024), p. 1367577.

[160] Ajay Dudani, Anjali Israni, Krish Dudani, and Anadya Dudani. "Revolutionizing cataract surgery with advanced techniques". In: *Indian Journal of Cataract and Refractive Surgery* 1.2 (2024), pp. 146–152.

[161] Marta Benítez Martínez, David Baeza Moyano, and Roberto Alonso González-Lezcano. "Phacoemulsification: Proposals for improvement in its application". In: *Healthcare*. Vol. 9. 11. MDPI. 2021, p. 1603.

[162] Balamurugan Ramatchandirane, Divya Sindhuja Pathuri, Mansi D Devalla, Shubhangi SN Prasad, Premika Pandarasamy, R Balamurugan, and Premika Pandarasamy Jr. "Two-Y crushing technique: A simple technique to crack the nucleus in a posterior polar cataract using Two-Y rotators". In: *Cureus* 16.6 (2024).

[163] Luoying Hao, Yan Hu, Wenjun Lin, Qun Wang, Heng Li, Huazhu Fu, Jinming Duan, and Jiang Liu. "Act-net: Anchor-context action detection in surgery videos". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 196–206.

[164] Abraham Tarekegn Mersh, Debas Yaregal Melesse, and Wubie Birlie Chekol. "A clinical perspective study on the compliance of surgical safety checklist in all surgical procedures done in operation theatres, in a teaching hospital, Ethiopia, 2021: A clinical perspective study". In: *Annals of Medicine and Surgery* 69 (2021), p. 102702.

[165] Dmitri Nepogodiev, Janet Martin, Bruce Biccard, Alex Makupe, Aneel Bhangu, Adesoji Ademuyiwa, Adewale Oluseye Adisa, Maria-Lorena Aguilera, Sohini Chakrabortee,

J Edward Fitzgerald, et al. "Global burden of postoperative death". In: *The Lancet* 393.10170 (2019), p. 401.

[166] Florent Lalys and Pierre Jannin. "Surgical process modelling: A review". In: *International Journal of Computer Assisted Radiology and Surgery* 9 (2014), pp. 495–511.

[167] Nicolas Padoy. "Machine and deep learning for workflow recognition during surgery". In: *Minimally Invasive Therapy & Allied Technologies* 28.2 (2019), pp. 82–90.

[168] D. A. Hashimoto, G. Rosman, E. R. Witkowski, C. Stafford, A. J. Navarette-Welton, D. W. Rattner, K. D. Lillemoe, D. L. Rus, and O. R. Meireles. "Computer vision analysis of intraoperative video: Automated recognition of operative steps in laparoscopic sleeve gastrectomy". In: *Annals of Surgery* (2019).

[169] Mobarakol Islam, Lalithkumar Seenivasan, Lim Chwee Ming, and Hongliang Ren. "Learning and reasoning with the graph structure representation in robotic surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 627–636.

[170] Jie Zhang, Shenchao Shi, Yiwei Wang, Chidan Wan, Huan Zhao, Xiong Cai, and Han Ding. "Automatic keyframe detection for critical actions from the experience of expert surgeons". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2022, pp. 8049–8056.

[171] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. "Leveraging uncertainty information from deep neural networks for disease detection". In: *Scientific Reports* 7.1 (2017), pp. 1–14.

[172] Michael P Linegang, Heather A Stoner, Michael J Patterson, Bobbie D Seppelt, Joshua D Hoffman, Zachariah B Crittendon, and John D Lee. "Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological ap-

proach". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 50. 23. SAGE Publications Sage CA: Los Angeles, CA. 2006, pp. 2482–2486.

[173] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems* 30 (2017).

[174] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1050–1059.

[175] Youngwan Lee, Joong-won Hwang, Hyung-Il Kim, Kimin Yun, Yongjin Kwon, Yuseok Bae, and Sung Ju Hwang. "Localization uncertainty estimation for anchor-free object detection". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2022, pp. 27–42.

[176] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10674–10685.

[177] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[178] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2016), pp. 1137–1149.

[179] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[180] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020).

[181] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. "Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents". In: *arXiv preprint arXiv:2201.00308* (2022).

[182] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[183] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. "Ssd: Single shot multibox detector". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2016, pp. 21–37.

[184] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017, pp. 2980–2988.

[185] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017, pp. 2961–2969.

[186] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4510–4520.

[187] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. "Dynamic r-cnn: Towards high quality object detection via dynamic training". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2020, pp. 260–275.

[188] Chengxin Liu, Kewei Wang, Hao Lu, Zhiguo Cao, and Ziming Zhang. "Robust object detection with inaccurate bounding boxes". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2022, pp. 53–69.

[189] Luoying Hao, Yan Hu, Yang Yue, Li Wu, Huazhu Fu, Jinming Duan, and Jiang Liu. "Hierarchical context transformer for multi-level semantic scene understanding". In: *IEEE Transactions on Circuits and Systems for Video Technology* 35.4 (2025), pp. 3134–3145.

[190] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. "SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 281–290.

[191] Zixiao Wang, Junwu Weng, Chun Yuan, and Jue Wang. "Truncate-split-contrast: a framework for learning from mislabeled videos". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 3. 2023, pp. 2751–2758.

[192] Arnaud Huaulmé, Pierre Jannin, Fabian Reche, Jean-Luc Faucheron, Alexandre Moreau-Gaudry, and Sandrine Voros. "Offline identification of surgical deviations in laparoscopic rectopexy". In: *Artificial Intelligence in Medicine* 104 (2020), p. 101837.

[193] Bokai Zhang, Bharti Goel, Mohammad Hasan Sarhan, Varun Kejriwal Goel, Rami Abukhalil, Bindu Kalesan, Natalie Stottler, and Svetlana Petculescu. "Surgical workflow recognition with temporal convolution and transformer for action segmentation". In: *International Journal of Computer Assisted Radiology and Surgery* 18.4 (2023), pp. 785–794.

[194] Pietro Mascagni, Armine Vardazaryan, Deepak Alapatt, Takeshi Urade, Taha Emre, Claudio Fiorillo, Patrick Pessaux, Didier Mutter, Jacques Marescaux, Guido Costamagna, et al. "Artificial intelligence for surgical safety: Automatic assessment of the

critical view of safety in laparoscopic cholecystectomy using deep learning". In: *Annals of Surgery* 275.5 (2022), pp. 955–961.

[195] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. "Masked motion encoding for self-supervised video representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2235–2245.

[196] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. "Towards unified surgical skill assessment". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 9522–9531.

[197] Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. "Hierarchical graph convolutional networks for action quality assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.12 (2023), pp. 7749–7763.

[198] Yuan-Ming Li, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. "Continual action assessment via task-consistent score-discriminative feature distribution modeling". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[199] Guanghui Yue, Di Cheng, Tianwei Zhou, Jingwen Hou, Weide Liu, Long Xu, Tianfu Wang, and Jun Cheng. "Perceptual quality assessment of enhanced colonoscopy images: A benchmark dataset and an objective method". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.10 (2023), pp. 5549–5561.

[200] Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. "Temporal memory relation network for workflow recognition from surgical video". In: *IEEE Transactions on Medical Imaging* 40.7 (2021), pp. 1911–1923.

[201] Fangqiu Yi, Yanfeng Yang, and Tingting Jiang. "Not end-to-end: Explore multi-stage architecture for online surgical phase recognition". In: *Proceedings of the Asian Conference on Computer Vision*. 2022, pp. 2613–2628.

[202] Xiaoying Pan, Xuanrong Gao, Hongyu Wang, Wuxia Zhang, Yuanzhen Mu, and Xianli He. "Temporal-based swin transformer network for workflow recognition of surgical video". In: *International Journal of Computer Assisted Radiology and Surgery* 18.1 (2023), pp. 139–147.

[203] Wenxi Yue, Hongen Liao, Yong Xia, Vincent Lam, Jiebo Luo, and Zhiyong Wang. "Cascade multi-level transformer network for surgical workflow analysis". In: *IEEE Transactions on Medical Imaging* 42.10 (2023), pp. 2817–2831.

[204] Nan Xi, Jingjing Meng, and Junsong Yuan. "Forest graph convolutional network for surgical action triplet recognition in endoscopic videos". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.12 (2022), pp. 8550–8561.

[205] Wei Wang, Xin Yang, and Jinhui Tang. "Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.9 (2023), pp. 4452–4461.

[206] Jiaxi Liu, Jinghao Niu, Weifeng Li, Xin Li, Binbin He, Hao Zhou, Yanjuan Liu, Ding Li, Bo Wang, and Wensheng Zhang. "Xfmp: A benchmark for explainable fine-grained abnormal behavior recognition on medical personal protective equipment". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[207] Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jue Wang, and Yuexian Zou. "Unsupervised pre-training for temporal action localization tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14031–14041.

[208] Ange Lou, Kareem Tawfik, Xing Yao, Ziteng Liu, and Jack Noble. "Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation". In: *IEEE Transactions on Medical Imaging* (2023).

[209] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. "St-adapter: Parameter-efficient image-to-video transfer learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26462–26477.

[210] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. "Aim: Adapting image models for efficient video action recognition". In: *The Eleventh International Conference on Learning Representations*. 2023.

[211] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. "Mvitv2: Improved multiscale vision transformers for classification and detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4804–4814.

[212] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. "Multiscale vision transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6824–6835.

[213] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. "Dino: Detr with improved denoising anchor boxes for end-to-end object detection". In: *The Eleventh International Conference on Learning Representations*. 2023.

[214] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. "Parameter-efficient transfer learning for NLP". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2790–2799.

[215] Sebastian Stein and Stephen 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*. 2013, pp. 729–738.

[216] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? A new model and the kinetics dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[217] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. "Slowfast networks for video recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6202–6211.

[218] Sepp Hochreiter and Jürgen Schmidhuber. "LSTM can solve hard long time lag problems". In: *Advances in Neural Information Processing Systems* 9 (1996).

[219] Yazan Abu Farha and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3575–3584.

[220] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. "Ms-tcn++: Multi-stage temporal convolutional network for action segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2020), pp. 6647–6658.

[221] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. "Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2022, pp. 52–68.

[222] Xiaoyan Tian, Ye Jin, and Xianglong Tang. "Local–global transformer neural network for temporal action segmentation". In: *Multimedia Systems* 29.2 (2023), pp. 615–626.

[223] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. "Asformer: Transformer for action segmentation". In: *arXiv preprint arXiv:2110.08568* (2021).

[224] Emad Bahrami, Gianpiero Francesca, and Juergen Gall. "How much temporal long-term context is needed for action segmentation?" In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 10351–10361.

[225] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. "Mediation analysis in epidemiology: Methods, interpretation and bias". In: *International Journal of Epidemiology* 42.5 (2013), pp. 1511–1519.

[226] Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. "Recursive causal structure learning in the presence of latent variables and selection bias". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10119–10130.

[227] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. "Interventional few-shot learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2734–2746.

[228] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. "Context de-confounded emotion recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19005–19015.

[229] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. "Two causal principles for improving visual dialog". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10860–10869.

[230] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. "Counterfactual vqa: A cause-effect look at language bias". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12700–12710.

[231] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. "Auto-parsing network for image captioning and visual question answering". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2197–2207.

[232] Kalanit Grill-Spector and Nancy Kanwisher. "Visual recognition: As soon as you know it is there, you know what it is". In: *Psychological Science* 16.2 (2005), pp. 152–160.

[233] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. "How to grow a mind: Statistics, structure, and abstraction". In: *Science* 331.6022 (2011), pp. 1279–1285.

[234] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. "Ms-tcn++: Multi-stage temporal convolutional network for action segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2020), pp. 6647–6658.

[235] Yueming Jin, Yonghao Long, Xiaojie Gao, Danail Stoyanov, Qi Dou, and Pheng-Ann Heng. "Trans-svnet: Hybrid embedding aggregation transformer for surgical workflow analysis". In: *International Journal of Computer Assisted Radiology and Surgery* 17.12 (2022), pp. 2193–2202.

[236] Rong Tao, Xiaoyang Zou, and Guoyan Zheng. "Last: Latent space-constrained transformers for automatic surgical phase recognition and tool presence detection". In: *IEEE Transactions on Medical Imaging* 42.11 (2023), pp. 3256–3268.

[237] Kun Yuan, Matthew Holden, Shijian Gao, and Wonsook Lee. "Anticipation for surgical workflow through instrument interaction and recognized signals". In: *Medical Image Analysis* 82 (2022), p. 102611.

[238] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[239] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[240] Luke Keele. "The statistics of causal inference: A view from political methodology". In: *Political Analysis* 23.3 (2015), pp. 313–335.

[241] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. "Deconfounding duration bias in watch-time prediction for video recommendation". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, pp. 4472–4481.

[242] Junyan Wang, Yiqi Jiang, Yang Long, Xiuyu Sun, Maurice Pagnucco, and Yang Song. "Deconfounding causal inference for zero-shot action recognition". In: *IEEE Transactions on Multimedia* (2023).

[243] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. "Visual commonsense r-cnn". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10760–10770.

[244] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2048–2057.

[245] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems* 30 (2017).

[246] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[247] Christoph Feichtenhofer. "X3d: Progressive network expansion for efficient video recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[248] Karttikeya Mangalam, Haoqi Fan, Yanghao Li, Chao-Yuan Wu, Bo Xiong, Christoph Feichtenhofer, and Jitendra Malik. "Reversible vision transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10830–10840.

[249] Rohit Girdhar and Kristen Grauman. "Anticipative video transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 13505–13515.

[250] Yang Liu, Maxence Boels, Luis C Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. "Lovit: Long video transformer for surgical phase recognition". In: *Medical Image Analysis* 99 (2025), p. 103366.

[251] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17 (2007), pp. 395–416.

[252] Lior Rokach and Oded Maimon. "Clustering methods". In: *Data Mining and Knowledge Discovery Handbook* (2005), pp. 321–352.

[253] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Knowledge Discovery and Data Mining*. Vol. 96. 34. 1996, pp. 226–231.

[254] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. "Finite mixture models". In: *Annual Review of Statistics and Its Application* 6.1 (2019), pp. 355–378.

[255] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. "Optics: Ordering points to identify the clustering structure". In: *ACM Sigmod Record* 28.2 (1999), pp. 49–60.

[256] David Arthur and Sergei Vassilvitskii. *K-means++: The advantages of careful seeding*. Tech. rep. Stanford, 2006.

[257] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. "The" something something" video database for learning and evaluating visual common sense". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017, pp. 5842–5850.

[258] Cong Li, Yongyan Fu, Shunming Liu, Honghua Yu, Xiaohong Yang, Meixia Zhang, and Lei Liu. "The global incidence and disability of eye injury: An analysis from the Global Burden of Disease Study 2019". In: *EClinicalMedicine* 62 (2023).

[259] Alfred Sommer, Hugh R Taylor, Thulasiraj D Ravilla, Sheila West, Thomas M Lietman, Jeremy D Keenan, Michael F Chiang, Alan L Robin, Richard P Mills, Council of the American Ophthalmological Society, et al. "Challenges of ophthalmic care in the developing world". In: *JAMA Ophthalmology* 132.5 (2014), pp. 640–644.

[260] Matthew J Burton, Jacqueline Ramke, Ana Patricia Marques, Rupert RA Bourne, Nathan Congdon, Iain Jones, Brandon AM Ah Tong, Simon Arunga, Damodar Bachani, Covadonga Bascaran, et al. "The lancet global health commission on global eye health: Vision beyond 2020". In: *The Lancet Global Health* 9.4 (2021), e489–e551.

[261] Kyle Lam and Jianing Qiu. "Foundation models: The future of surgical artificial intelligence?" In: *British Journal of Surgery* 111.4 (2024), znae090.

[262] Hani J Marcus, Pedro T Ramirez, Danyal Z Khan, Hugo Layard Horsfall, John G Hanrahan, Simon C Williams, David J Beard, Rani Bhat, Ken Catchpole, Andrew

Cook, et al. "The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring". In: *Nature Medicine* 30.1 (2024), pp. 61–75.

[263] TL Edwards, K Xue, HCM Meenink, MJ Beelen, GJL Naus, MP Simunovic, M Latasiewicz, AD Farmery, MD De Smet, and RE MacLaren. "First-in-human study of the safety and viability of intraocular robotic surgery". In: *Nature Biomedical Engineering* 2.9 (2018), pp. 649–656.

[264] Dani Kiyasseh, Runzhuo Ma, Taseen F Haque, Brian J Miles, Christian Wagner, Daniel A Donoho, Animashree Anandkumar, and Andrew J Hung. "A vision transformer for decoding surgeon activity from surgical videos". In: *Nature Biomedical Engineering* 7.6 (2023), pp. 780–796.

[265] Jianfeng Cao, Hon-Chi Yip, Yueyao Chen, Markus Scheppach, Xiaobei Luo, Hongzheng Yang, Ming Kit Cheng, Yonghao Long, Yueming Jin, Philip Wai-Yan Chiu, et al. "Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study". In: *Nature Communications* 14.1 (2023), p. 6676.

[266] Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. "Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2024, pp. 481–500.

[267] Siddhant Satyanaik, Aditya Murali, Deepak Alapatt, Xin Wang, Pietro Mascagni, and Nicolas Padoy. "Optimizing latent graph representations of surgical scenes for unseen domain generalization". In: *International Journal of Computer Assisted Radiology and Surgery* 19.6 (2024), pp. 1243–1250.

[268] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An image is worth 16x16 words:

Transformers for image recognition at scale". In: *International Conference on Learning Representations*. 2021.

[269] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. "Efficientsam: Leveraged masked image pretraining for efficient segment anything". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 16111–16121.

[270] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. "Tinyvit: Fast pretraining distillation for small vision transformers". In: *Proceedings of the European Conference on Computer Vision*. Springer. 2022, pp. 68–85.

[271] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. "Cataract-101: Video dataset of 101 cataract surgeries". In: *Proceedings of the 9th ACM Multimedia Systems Conference*. 2018, pp. 421–425.

[272] Imanol Luengo, Maria Grammatikopoulou, Rahim Mohammadi, Chris Walsh, Chinedu Innocent Nwoye, Deepak Alapatt, Nicolas Padoy, Zhen-Liang Ni, Chen-Chen Fan, Gui-Bin Bian, et al. "2020 cataracts semantic segmentation challenge". In: *arXiv preprint arXiv:2110.10965* (2021).

[273] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159.

[274] Ga-In Lee, Dong Hui Lim, Sang Ah Chi, Seon Woo Kim, Dong Wook Shin, and Tae-Young Chung. "Risk factors for intraocular lens dislocation after phacoemulsification: a nationwide population-based cohort study". In: *American Journal of Ophthalmology* 214 (2020), pp. 86–96.

[275] Erkan Ünsal, Kadir Eltutar, and İlkay K Muftuoglu. "Morphologic changes in the anterior segment using ultrasound biomicroscopy after cataract surgery and intraocular lens implantation". In: *European Journal of Ophthalmology* 27.1 (2017), pp. 31–38.