

KYOTO UNIVERSITY

**Intraoperative Situational Understanding
from Operating Room Surveillance Videos
Using Human Activity Recognition**

by

Koji Yokoyama

A dissertation submitted in partial fulfillment for the
degree of Doctor of Philosophy
in

Medical Informatics
Department of Social Informatics
Graduate School of Informatics

March 2025

KYOTO UNIVERSITY

Abstract

Medical Informatics
Department of Social Informatics
Graduate School of Informatics

Doctor of Philosophy

by Koji Yokoyama

Since unexpected incidents in the operating room can significantly impact a patient’s life, appropriate medical safety measures must be considered. Modern operating rooms are complex environments where human errors such as inattention, miscommunication, insufficient training, and improper equipment use are cited as leading causes of accidents. However, current safety measures rely on subjective records, such as written notes and electronic logs, without objectively analyzing the activities of surgical team members. To address this limitation, examination of video footage of events in the operating room is necessary.

Recently, the installation of surveillance cameras that capture the entire operating room has increased, enabling the recording of surgeries. By extracting features of human activity from these videos and analyzing the variations based on surgical progression, emergencies, and workload, we aim to enhance surgical safety and operational efficiency through objective event detection and situation awareness.

First, we propose two methods for recognizing group activities specific to surgeries: a) medical instrument passing and b) the attention activity of surgical team members. The medical instruments passing among nurses corresponds to the progression of the surgery, and deviations in the type and timing of instrument handovers may procedural irregularities, such as emergencies. Similarly, recognizing the attention activity of surgical team members provides insights into incident detections, as surgeons and scrub nurses tend to focus on the vital signs monitor during critical events. To achieve this, we extracted features for individual activity recognition including body orientation, face orientation, and arm extension, using pose estimation from videos. Medical instrument passing was then recognized using a (Long Short-Term Memory) LSTM-based deep learning model for time-series analysis. Additionally, we generated heat maps of attention areas based on the face direction of surgical team members to estimate focal points during surgery.

Second, we sought to extract the states and features of individual activities. Since positions and movements in the operating room are role-specific, detecting deviations from expected activity patterns could facilitate event detection and situational recognition. To achieve this we proposed three models: a) activity anomaly estimation, b) activity clustering, and c) role classification.

For a), we developed a model combining an autoencoder and a generative adversarial network to assess the degree of activity anomalies based on reconstruction errors over five-second intervals. Since the positions and postures of surgical team members remain relatively stable during smooth surgical progression, training the model exclusively on these stable activities allows it to assign high anomaly scores to individuals displaying significant movement. We confirmed that high anomaly scores were observed during

uncommon activities, such as circulators moving around to retrieve medical instruments or crouching to measure blood loss with gauze.

For b), we employed unsupervised learning to extract and model activity features for each role in the operating room. Changes in these activity patterns could be leveraged for incident detection, workload assessment, and overall surgical safety and efficiency improvements. However, distinguishing between different surgical team members' movements is challenging due to their largely stationary nature, necessitating annotations by experienced practitioners, which increases annotation costs. Surgery in the operating room involves surgical team members positioned around the operating table. To leverage this feature, we manually specified the center coordinates of the operating table and the L2 norm of each individual's bounding box relative to this center as an input feature. The results demonstrated that activity classification achieved over 70% accuracy on the training dataset and over 60% accuracy on the test dataset. Further incorporating spatial features improved the accuracy by 10% leading to final classification accuracies of approximately 82.6% for the training dataset and 70.6% for the test dataset.

For c), we identified surgical team members' roles (e.g., surgeon, anesthetist, scrub nurse, circulator, others) using self-supervised learning. To address the challenge of applying supervised learning models, due to the difficulty of obtaining accurate annotations and achieving robust learning from surgical videos, we proposed a self-supervised learning model using a categorical variational autoencoder. Using a dataset derived from 1 hour of video, this approach achieved approximately 95% accuracy in role identification with annotations for only 1/4 of the total data.

Using these activity recognition results, we conducted event detection and situational recognition to enhance surgical safety and efficiency. Our attention activity analysis successfully detected events such as surgeons leaving the surgical field to gather in front of the monitor to confer. Additionally, role classification results enabled the detection of emergencies, such as massive bleeding (cardiac arrest), based on sudden changes in the number of personnel in the operating room. These findings demonstrate the potential for using surgical video analysis to detect critical events and improve surgical safety and workflow efficiency.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Professor Tomohiro Kuroda for accepting to be my supervisor. His unwavering support, invaluable advice, and expert guidance have been instrumental in conducting this research. His kindness and willingness to share his vast knowledge have made a significant impact on my academic journey.

With equal emotion and sincerity, I extend my sincere gratitude to Professor Goshiro Yamamoto, who has not only imparted his knowledge and experience as a teacher but also offered unwavering support and encouragement. His mentorship has been invaluable, and his dedication to walking side by side with me on this challenging path has provided immense motivation and inspiration.

I am profoundly thankful to Professors Chang Liu, Sho Mitarai, Kazumasa Kishimoto and Yukiko Mori, who have collaborated with our team from the very beginning of this research. Their tireless efforts and contributions made all the experiments and evaluations of this work possible. They provided not only essential datasets but also critical insights from a doctor's perspective, which have been pivotal in shaping the research outcomes.

I am also grateful to Associate Professor Kei Hiroi of the Department of Disaster Prevention Research Institute, Department of Social Informatics, Graduate School of Informatics, Kyoto University, and Doctor Hideya Seo, Kyoto University Hospital for their willingness to serve as advisors despite their demanding schedules. Their invaluable guidance, thoughtful advice, and constructive feedback have been indispensable in refining and advancing this research.

Equally, I extend my heartfelt thanks to all the staff of Medical Informatics Lab, including Professors Kazuya Okamoto, Osamu Sugiyama, Shusuke Hiragi, Luciano HO Santos, Hiroshi Sasaki, Hiroki Ueshima, and Kenichi Saito. Their valuable comments, diverse perspectives, and intellectual contributions have greatly enriched this research. Their encouragement to consider multiple viewpoints has deepened the discussions and led to more fruitful outcomes.

I would also like to express my gratitude to the students Natsuko Nishida and Akihito Ito. Their advice, knowledge-sharing, and contributions have been immensely helpful.

A very special thank you goes to the hardworking Medical Informatics secretaries, Yoko Hara, Yuko Furusawa, and Yuki Ozawa, without whom no one in this lab, be they students or professors, would be able to accomplish anything. Their meticulous assistance in managing numerous documents and deadlines, as well as their timely reminders

and scoldings, have been vital to our productivity and success. Their dedication and hard work are truly appreciated.

February 25th, 2025

Koji Yokoyama

Ethical Approval

This research was approved by the Ethics Committee of the Graduate School and Faculty of Medicine, Kyoto University, R3282.

Contents

Abstract	i
Acknowledgements	iv
Ethical Approval	vi
List of Figures	xi
List of Tables	xvii
1 Introduction	1
1.1 Challenges of Surgery	1
1.2 Subjective Surgical Analysis Methods	1
1.3 Reasons for Selecting Operating Room Surveillance Video	2
1.4 Related Work – Vision-Based HAR	2
1.5 Surgical Team Members Captured in Operating Room Surveillance Video	5
1.6 Applications of Vision-Based HAR for Surgery	6
1.7 Research Scope and Conditions	7
1.8 Dissertation Overview	8
I Linear Transformation-Based Group Activity Recognition	10
2 Linear Transformation-Based Individual Activity Features	12
2.1 Introduction	12
2.2 Data Pipeline of Group Activity Recognition	13
2.3 Notations	14
2.4 Pose Estimation	14
2.5 Tracking Using Particle Filter	14
2.6 Individual Activity Features	15
2.6.1 Body Position	15
2.6.2 Face Orientation and Body Orientation	15
2.6.3 Arm Flexion	16
2.6.4 Wrist Position	17
2.7 Summary of Individual Activity Features	17
3 Medical Instrument Passing	18
3.1 Introduction	18
3.2 Methods	18

3.2.1	Extracting Features	18
3.2.2	Machine-Learning-Based Models	20
3.3	Experiments	20
3.3.1	Simulated Videos	20
3.3.2	Real Operating Room Surveillance Videos	23
3.4	Discussion	27
3.5	Conclusion	30
4	Group Attention	31
4.1	Introduction	31
4.2	Methods	31
4.3	Experiments	33
4.3.1	Simulated Videos	33
4.3.2	Real Operating Room Surveillance Videos	36
4.4	Discussion	39
4.5	Conclusion	43
II	End-to-End Individual Activity Recognition	44
5	Unsupervised Anomaly Estimation	46
5.1	Introduction	46
5.2	Related Work	47
5.2.1	Reconstruction-Based Human Activity Anomaly Detection	47
5.2.2	Our Motivation	47
5.3	Methods	48
5.3.1	Pose Estimation and Tracking	48
5.3.2	Activity Anomaly Estimation using Time-Sequential Prediction	49
5.4	Experiments	49
5.4.1	Dataset	49
5.4.2	Implementation	50
5.4.3	Results	50
5.5	Discussion	51
5.6	Conclusion	52
6	Unsupervised Activity Clustering	53
6.1	Introduction	53
6.2	Related Work	54
6.2.1	Deep Clustering	54
6.2.2	Our Motivation	54
6.3	Methods	55
6.3.1	Model Architecture	55
6.3.2	Clustering Algorithm	57
6.3.3	Loss Functions	58
6.3.4	Training Algorithm	58
6.4	Experiments and Results	58
6.4.1	Dataset	58
6.4.2	Implementation	61

6.4.3	Results	65
6.5	Discussion	67
6.5.1	Implication	67
6.5.2	Effect of Spatial Feature	68
6.5.3	Limitation	68
6.6	Conclusion	68
7	Semi-Supervised Role Classification	70
7.1	Introduction	70
7.2	Theoretical Background	71
7.2.1	VAE	71
7.2.2	Categorical VAE	72
7.2.3	VQ-VAE	73
7.2.4	SQ-VAE	73
7.3	Categorical SQ-VAE	75
7.4	Experiments	78
7.4.1	Model Architecture	78
7.4.2	Data Sampling	79
7.4.3	Scaling	80
7.4.4	Role Labels	80
7.4.5	Annotation Protocol	80
7.4.6	Contents of Dataset	82
7.4.7	Training Algorithm	82
7.4.8	Results	83
7.5	Discussion	87
7.6	Conclusion	92
III	Intraoperative Situational Analysis	94
8	Intraoperative Situation Recognition and Event Detection	95
8.1	Introduction	95
8.2	Linear Transformation-Based Group Activity Recognition	96
8.2.1	Methods	96
8.2.2	Dataset	96
8.2.3	Results	99
8.3	End-to-End Individual Activity Recognition	103
8.3.1	Methods	103
8.3.2	Dataset	103
8.3.3	Prediction	105
8.3.4	Results	105
8.4	Discussion	114
8.5	Conclusion	114
9	Conclusion	115
9.1	Contributions	115
9.2	Future Direction	119
9.2.1	Required Research Advancements for Surgical Application	119

9.2.2	Effective Utilization of Operating Room Surveillance Video	120
9.2.3	Research Environments	121
9.2.4	Multimodel Model	122
References		123
List of Publications		135

List of Figures

1.1	(a) Overview of the operating room. The surgeons, anesthetist, and scrub nurse remain in fixed positions, while the circulator moves around the operating room to deliver medical instruments. (b) A sample frame from an actual operating room surveillance video.	4
1.2	The research framework of vision-based HAR for operating room surveillance videos.	7
1.3	This dissertation comprises three main parts: a) Linear Transformation-Based Group Activity Recognition, b) End-to-End Individual Activity Recognition, and c) Intraoperative Analysis.	9
2.1	Data pipeline of group activity recognition comprising three main steps: 1) pose estimation, 2) individual activity feature extraction (Chapter 2), and 3) group activity recognition. We present two group activity recognition models: a) medical instrument passing (Chapter 3) and b) group attention (Chapter 4).	13
3.1	Features of MIP.	20
3.2	Architecture of RNN using LSTM. The model was composed of an LSTM layer, two linear layers, and a log-softmax layer. The features of the last T frames were input into the RNN model. The model yielded a 2D output that included the likelihoods of passing and not passing medical instrument.	21
3.3	Images showing (a) the positions for MIP IP1 (orange) and IP2 (violet), and (b) the actual video frame of IP1	22
3.4	Results of the SVM model for predicting instrument passing activities in comparison with the ground truth. The solid blue line represents the prediction result and the dashed orange line represents the ground truth. Other lines represent other features in terms of “distance”: the Euclidean distance of position between the two individuals; “body”: the body orientation two individuals in relation to each other; “arm”: the average arm flexion; and “wrist”: the minimum Euclidean distance between the wrists of two individuals.	24
3.5	Sample results of one case of MIP activities the SVM model could not predict while RNN model could. The solid blue line represents the prediction result and the dashed orange line represents the ground truth. Other lines represent features in terms of “distance”: the Euclidean distance of position between the two individuals; “body”: the body orientation between two individuals; “arm”: the average arm flexion; and “wrist”: the minimum Euclidean distance between the wrists of two individuals.	27

3.6	Samples of actual frames of operating room surveillance videos. For the top, MIP was marked between the circulator (Individual 1) and the scrub nurse (Individual 3). At the bottom, MIP was marked between the circulator (Individual 2) and scrub nurse (individual 4).	28
3.7	Scatter plots of MIP features S , Left: simulated videos; Right: actual operating room surveillance videos. Red and blue points are MIP activity features and Non-MIP activity features, respectively.	29
3.8	Sample failure result of estimation of operating room surveillance videos. The circulator was occluded by the monitor and was, therefore, undetected by pose estimation.	30
4.1	Diagram of viewing field of individual I_i at Frame t . A point X in the example in this figure, $ \theta_{i X} \leq \phi$, and $l_{i X} > \mu$, where $\phi = 10$ deg and $\mu = 150$ px. Thus, $p_{i,t X}$ was calculated using the middle case of Equation 4.1.	32
4.2	Diagram showing estimating attention score $s_{n T}$ of O_n at frame T . In the example shown in this figure, there were three individuals, I_1, I_2 , and I_3 . The attention probability of $p_{T X} = 3$ because X was in the view field of all three individuals.	32
4.3	Images (a) showing the positions for group attention GA1 (orange), GA2 (violet), and GA3 (light blue), and (b) the actual frame of GA1	34
4.4	Histograms of attention score results. The top, middle, and last rows show the results for GA1 , GA2 , and GA3 , respectively. The legend of " O_n -N/A" denotes that the actors were not paying attention to O_n and " O_n -A" paying attention to O_n , where $n = 1, 2, 3$	35
4.5	Actual video frames in last 10s of GA1 . The top panel shows that the actors were looking at O_1 (monitor). The middle shows that the actors were looking at O_2 (PC). The bottom panel shows that the actors were looking at O_3 (Operating Table).	37
4.6	Placement of objects of focus in actual operating room surveillance videos. (a) The objects in S1 were O_1 , a monitor under the surveillance camera; O_2 , a dissection table where STs were carried out following the excision of the affected part; and O_3 , the operating table. (b) The objects in S2 were O_1 , a PC placed to right-back from the surveillance camera; O_2 , the dissection table where STs were performed after excision of the affected part; O_3 , the operating table. (c) The objects in S3 were O_1 , a monitor under the surveillance camera; O_2 , a PC placed near the monitor; and O_3 , the operating table.	38
4.7	Histograms of attention score results. The top, middle, and bottom rows show the results for S1 , S2 , and S3 , respectively. The legend of " O_n -N/A" denotes that the surgical team members were not paying attention to O_n and " O_n -A" denotes that they were paying attention to O_n	40
4.8	Actual video frames of S1 - O_1 . The top shows the frame where all team members were not looking at the O_1 (monitor) and the bottom row shows the frames in which some team members gathered around the O_1 (monitor), gazing at it.	41

4.9	(Top) Image showing boxplots of the confidence scores of the pose estimation of the nose, left and right ear points in the simulated videos. (Bottom) The actual frames of GA1 - O₁ and GA1 - O₂ . For GA1 - O₁ , because the actors were facing the surveillance camera, the confidence scores were high. However, for GA1 - O₂ , the confidence score of the right ear was low because the actors turned sideways from the surveillance camera.	42
4.10	Wrong face orientation sample. The face orientation of Individual 1 (red arrow) was turned left. However, the expected face orientation is indicated by an orange arrow.	42
5.1	System flow comprising two components: 1) pose estimation and tracking, and 2) GAN-based individual activity recognition.	48
5.2	The architecture of GAN-based individual activity recognition model which consists of (a) generator and (b) discriminator.	48
5.3	The result samples of anomaly score: (a) training dataset and (b) test dataset. The top of each result was estimated by bounding boxes. The bottom of each result was estimated by key points. The labels are tracking IDs. The red arrows show the situation of Figure 5.4 (a) and (b).	51
5.4	The real data and fake data indicated by the red arrows in Figure 5.3: (a) training datasets and (b) test datasets. Green and red lines show the real data and the fake data, respectively. The top of each shows the top-left corner of bounding box and the bottom of each shows key points at 30 frames (1 sec) intervals.	51
6.1	Overview of the proposed model. Visual features are extracted from two autoencoders (RGB frame, optical flow). In the clustering layer, individual features are extracted from the given bounding boxes, and spatial features from the center of the operating table and the center are added. Finally, the cluster number of each individual is output.	56
6.2	Model architectures of (a) Decoders and (b) Clustering Layers.	56
6.3	Sample frames of operating room videos. (a) to (f) corresponding in order from <i>S1</i> to <i>S6</i> . Each red point is manually pointed to the center of operating table in each surgery.	60
6.4	Label samples of the dataset that includes surgeon, anesthetist, scrub nurse, circulator, and other.	60
6.5	Example of linking clustering results to true labels. Cluster 1 (orange) was linked to circulator, cluster 2 (green) was linked to surgeon, and cluster 3 (red) was linked to surgeon.	62
6.6	Confusion matrix of the training dataset. (a) to (d) correspond in order from condition 1 to 4 of Table 6.3. The labels are set as follows; 0: surgeon, 1: anesthetist, 2: scrub nurse, 3: circulator, 4: other.	63
6.7	Confusion matrix of the test dataset. (a) to (d) correspond in order from condition 1 to 4 of Table 6.3. The labels are set as follows; 0: surgeon, 1: anesthetist, 2: scrub nurse, 3: circulator, 4: other.	64
6.8	Clustering result samples of the training dataset. (a) to (d) correspond in order from condition 1 to 4 of Table 6.3. The true label and predicted label are shown in each white text box.	66
6.9	Clustering result of the test dataset using the model of condition 4 of Table 6.3. (a) to (f) correspond in order from <i>S1</i> to <i>S6</i>	66

7.1	Overview of Categorical SQ-VAE, which consists of Encoder, Decoder, and Classification Head, and the latent variable spaces for each label.	75
7.2	Overview of the proposed model architecture based on Figure 7.1. From left to right, the encoder $q_\phi(z x)$, classification head $q_\phi(c z)$, and decoder $p_\theta(x z_q)$ are arranged.	78
7.3	Sample frames of operating room surveillance videos. The frames from Surgery 1 to Surgery 6 are arranged in order from (a) to (f).	81
7.4	Confusion matrices of training dataset. The results for S1 to S3 under training condition T1 are arranged in order from (a) to (c). The result for S1 under condition T2 in (d), for S2 under condition T3 in (e), and for S3 under condition T4 in (f).	84
7.5	Confusion matrices of the test dataset. The results from S1 to S6 under training condition T1 are arranged in order from (a) to (f). The red-framed results were used for training.	85
7.6	Confusion matrices of the test dataset. The results from S1 to S6 under training condition T2 are arranged in order from (a) to (f). The red-framed result was used for training.	85
7.7	Confusion matrices of the test dataset. The results from S1 to S6 under training condition T3 are arranged in order from (a) to (f). The red-framed result was used for training.	86
7.8	Confusion matrices of the test dataset. The results from S1 to S6 under training condition T4 are arranged in order from (a) to (f). The red-framed result was used for training.	86
7.9	Box plots of MSE for keypoints and bounding boxes under each learning condition. (a) to (d) show the results for keypoints, and (e) to (h) show the results for bounding boxes. The red diamonds indicate that the surgeries were used for training.	87
7.10	Scatter plot of bounding boxes for misclassified samples from S1 to S3. The left column shows bounding boxes colored by the ground truth, while the right column shows bounding boxes colored by the falsely predicted results under training condition T1.	88
7.11	Scatter plot of bounding boxes for misclassified samples from S4 to S6. The left column shows bounding boxes colored by the ground truth, while the right column shows bounding boxes colored by the falsely predicted results under training condition T1.	89
7.12	Correctly classified samples (left) and misclassified samples (right) for circulator (label 3).	90
7.13	Correctly classified samples (left) misclassified samples (right) of other (label 4).	91
8.1	Results of group activity recognition for Surgery 1 (top), Surgery 2 (middle), and Surgery 3 (bottom).	97
8.2	Results of group activity recognition for Surgery 4 (Top), Surgery 5 (Middle), and Surgery 6 (Bottom).	98
8.3	Examples of activities recognized by the MIP model involving two surgical team members working together. (Top)Two surgical team members cover a patient with a blanket. (Bottom)Two surgical team members dissect the excised tumor.	99

8.4	Frames showing the draping process in each operating room surveillance video (“Start” in Figure 8.1 and Figure 8.2). The surgeons performing the draping are highlighted with red boxes. Results from surgeries 1 to 6 are arranged in order from (a) to (f).	100
8.5	Frames showing the draping removal process in each operating room surveillance video (“End” in Figure 8.1 and Figure 8.2). The surgeons removing the draping are highlighted with red boxes. Results from surgeries 1 to 6 are arranged in order from (a) to (f).	100
8.6	Examples of discrepancies between AP movements and the number of surgical team members. (Top) Results around the 3 hours mark of Surgery 1. Interns (ID:3827 and 3947) are conversing in front of the monitor. (Middle) Results around the 3 hours 7 minutes mark of Surgery 5. Interns (ID:1460, 1576, 1597, and 1642) are conversing in a corner of the operating room. (Bottom) Results around the 3 hours 15 minutes mark of Surgery 6. Interns (ID:2603, 2652, 2653, and 2656) are conversing in a corner of the operating room.	101
8.7	Clustering the latent variables of the CSQ-VAE using K-means (k=5) for each role. The latent variable z is input into the K-means model for classification.	104
8.8	Changes in the number of predicted roles and event detection results for Surgery 1. The videos were segmented into 30 minutes intervals, capturing the start and end of surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3.	106
8.9	Changes in the number of predicted roles and event detection results for Surgery 2. The videos were segmented into 30 minutes intervals, capturing the start and end of surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3.	107
8.10	Changes in the number of predicted roles and event detection results for Surgery 3. The videos were segmented into 30 minutes intervals, capturing the start and end of surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3.	108
8.11	Sample frames capturing event occurrence during Surgery 1.	109
8.12	Sample frames capturing event occurrence during Surgery 2.	109
8.13	Sample frames capturing event occurrence during Surgery 3.	109
8.14	Changes in the number of K-means clustering results for Surgery 1. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The “period” represents the time elapsed within the entire surgery.	110

8.15	Changes in the number of K-means clustering results for Surgery 2. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The "period" represents the time elapsed within the entire surgery.	111
8.16	Changes in the number of K-means clustering results for Surgery 3. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The "period" represents the time elapsed within the entire surgery.	112
9.1	Overview of future directions.	120

List of Tables

3.1	Composition of datasets by frame number of MIP activities.	22
3.2	Prediction results of the MIP activity.	23
3.3	Summary of MIP data. The number of MIPs is the number of times MIP activities performed in each surgery.	24
3.4	Composition of datasets by number of MIP activities. The training data had 50 MIP activities and 5583 Non-MIP activities. The test data included 18 MIP activities and 2393 Non-MIP activities. The ratio of the total training data to the total test data was approximately 7 : 3. Both the training data and test data contained significantly more Non-MIP activity because there were many combinations of STs not performing MIP.	25
3.5	Composition of training datasets by number of frame of MIP activities in actual surgical video for learning SVM model.	25
3.6	Prediction scores of the MIP activity of test datasets per frame.	26
3.7	Prediction scores of the MIP activity of test dataset per event.	26
4.1	P-values of one-sided Mann-Whitney U comparison test of the attention scores of $\mathbf{O}_n\text{-A}$ and those of $\mathbf{O}_n\text{-N/A}$ of GA1 , GA2 , and GA3 . The null hypothesis was set at $\mu_n\text{-A} = \mu_n\text{-N/A}$, and the alternative hypothesis was set at $\mu_n\text{-A} > \mu_n\text{-N/A}$	36
4.2	Summary of GA data. The objects of focus in each surgery are indicated. The GA frame is the total number of frames for the group attention activities. The Non-GA frame is the total number of frames where the participants did not pay attention to any particular object. Each video was cropped to be one-minute long ($30 \text{ FPS} \times 60\text{s} = 1800 \text{ frames}$).	38
4.3	P-values of one-sided Mann-Whitney U test for the comparison of the attention scores of $\mathbf{O}_n\text{-A}$ and those of $\mathbf{O}_n\text{-N/A}$ of S1 , S2 , and S3 . The null and alternative hypotheses were set as $\mu_n\text{-A} = \mu_n\text{-N/A}$ and $\mu_n\text{-A} > \mu_n\text{-N/A}$, respectively.	39
6.1	Dataset contents. Training dataset consists of 20 videos from three surgeries $S1 \cdots S3$. Test dataset consists of nine videos from six surgeries $S1 \cdots S6$ (three of videos are the same surgeries of training dataset).	61
6.2	Structure of 3D transposed convolution layers in the decoder.	61
6.3	Conditions of training.	61
7.1	The architecture of the Embedding layer.	79
7.2	The architecture of the Reconstruction layer.	79
7.3	The number of samples of training dataset.	81
7.4	The number of samples of test dataset.	81

7.5	The training conditions. The checkmarks indicate which datasets were used for each training condition.	83
7.6	The accuracy scores of training dataset.	83
7.7	The accuracy scores of test dataset.	83
8.1	Contents of operating room surveillance videos used for event detection and situation recognition with handcrafted feature-based group activity recognition.	99
8.2	Contents of operating room surveillance videos used for event detection and situational recognition through end-to-end individual activity recognition.	104

Chapter 1

Introduction

1.1 Challenges of Surgery

Medical institutions have inevitably tightened their safety protocols in response to medical incidents [1]. Surgery requires rigorous safety measures to prevent unexpected complications during procedures. The operating room is a highly complex environment, making surgical procedures particularly vulnerable to errors caused by lapses in attention among team members [2]. Currently, approximately 312.9 million surgeries are performed worldwide [3]. A medical field study reported that 7,508 out of 44,814 adult patients (16.8%) developed at least one postoperative complication, with 207 (2.8%) dying as a result [4]. Additionally, 45% of all medical accidents in healthcare occur in the operating room, and over 50% of postoperative complications are attributed to human error. These errors stem from factors such as poor teamwork, inadequate communication, hierarchical barriers, inexperience, lack of training, improper equipment use, and carelessness [5–9]. A major challenge in improving surgical safety is identifying the causes of intraoperative human errors and preventing medical accidents. Addressing this challenge requires effective surgical analysis methods.

1.2 Subjective Surgical Analysis Methods

Several methods for improving surgical safety rely on subjective evaluation. Current approaches to assessing human error in surgery include evaluating critical surgical steps and teamwork through checklists and postoperative interviews. The Observational Teamwork Assessment for Surgery (OTAS) [10] evaluates teamwork-related activities using a 15-item checklist on a 7-point scale. The Operating Theater Team Non-Technical Skills

Assessment Tool (NOTECHS II) [11] employs an 8-point scale to assess various dimensions of non-technical skills, providing higher accuracy and greater sensitivity in surgical evaluations [12]. Additionally, the Nontechnical Skills for Surgeons (NOTSS) [13] assesses five key competencies, situation recognition, decision-making, communication, teamwork, and leadership, using a 4-point scale. A 2015 review of 25 studies concluded that NOTSS offers the highest validity, reliability, and acceptability among these methods [14].

However, these evaluation methods have limitations. They assess only predefined criteria and rely on the subjective judgments of surgical team members, making it difficult to evaluate unanticipated factors or elements beyond the team's awareness. Furthermore, they do not directly analyze activities at the moment human errors occur, nor do they specify causal events. Therefore, it is necessary to extract and analyze team members' intraoperative activities from video footage and other objective data sources to systematically identify the causes of human error.

1.3 Reasons for Selecting Operating Room Surveillance Video

Intraoperative records maintained by surgical staff often rely on handwritten notes, making real-time documentation challenging. During emergencies, staff are too occupied to record detailed events, resulting in a lack of accurate and comprehensive intraoperative records suitable for surgical analysis.

This study focuses on operating room surveillance videos. With advancements in video recording technology, surveillance cameras capable of capturing the entire operating room are increasingly being installed. The use of these cameras offers a significant advantage: they eliminate the need for additional sensors while enabling the development of safety-enhancing technologies without interfering with surgical procedures.

This study uses vision-based human activity recognition (HAR) to automatically extract and analyze intraoperative situations from operating room surveillance videos, providing an objective approach to identify human errors and improve surgical safety.

1.4 Related Work – Vision-Based HAR

HAR is one of the most prominent fields in computer vision that focuses on analyzing human movements, posture control, and interactions. Using sensors embedded in

wearable devices and smartphones, such as microphones, cameras, accelerometers, and gyroscopes, various types of data can be measured to recognize human activities based on body movements and positions [15].

Advancements in video technology have made it easier to record videos anytime and anywhere using smartphones and surveillance cameras. For example, video recording is now commonplace at construction sites, factories, retail stores, households, in dashboard cameras, and during sports activities. Human activities captured in these videos are being actively researched for applications across various fields and levels of complexity. These include safety monitoring, environmental surveillance [16, 17], robotics [18], real-time training feedback [19], healthcare, medical diagnostics, fitness monitoring [20], and biomechanical analysis using data-driven approaches [21]. Extracting meaningful information from such videos through activity recognition is an ongoing area of research.

Group activity recognition (GAR), a form of HAR involving multiple people in the video, has been extensively studied. With the emergence of deep learning, numerous models have been proposed for HAR, surpassing traditional hand-crafted feature-based approaches. Early frameworks focused on modeling temporal dynamics using recurrent neural networks (RNNs) [22–24]. Ibrahim et al. [22] and Kong et al. [24] introduced models utilizing long short-term memory (LSTM) layers to construct hierarchical networks, extract temporal features, and focus on key individuals and their contexts. Wang et al. [23] proposed a recurrent context model that considers individual, group, and scene-level interactions.

Spatio-temporal relations are critical for enhancing HAR performance [25–31]. Qi et al. [25] constructed semantic graphs to describe spatio-temporal contexts by passing messages through structural RNN nodes. Azar et al. [26] developed a convolutional relational model that learns intermediate person-relationship representations based on video features extracted using an Inception 3D (I3D) backbone [32]. Wu et al. [27] employed two-dimensional (2D) convolutional neural networks (CNNs) and graph CNNs to construct individual relationship graphs, analyzing appearance and location-based relationships between individuals. Yuan et al. [28] introduced a dynamic graph for modeling diverse person interactions. Yan et al. [29] proposed a hierarchical graph-based cross-inference network to capture activity at the body-region, individual, and group levels. Li et al. [30] designed a panoramic graph incorporating multi-person skeletons and objects, using a Graph Convolutional Network (GCN) to model intra-person, inter-person, and person-object interactions. Jiang et al. [31] extracted both local positional features of individual actors and global features of the overall scene, using a graph neural network to infer interactions between actors and their environment.

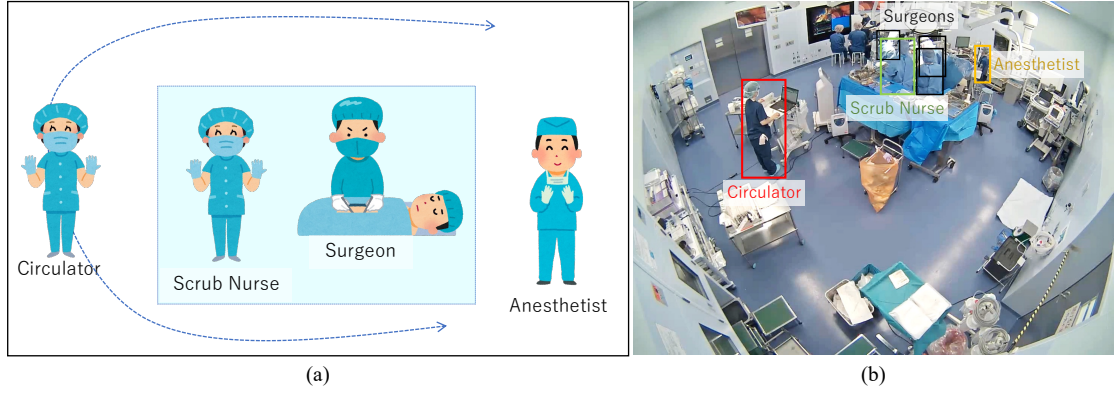


Figure 1.1: (a) Overview of the operating room. The surgeons, anesthetist, and scrub nurse remain in fixed positions, while the circulator moves around the operating room to deliver medical instruments. (b) A sample frame from an actual operating room surveillance video.

Inspired by the success of transformers in natural language processing (NLP) and computer vision, novel architectures integrating transformers for HAR have been proposed [33–36]. Gavriluk et al. [33] introduced an actor-transformer model to selectively extract information relevant to group activity recognition. Yuan et al. [34] developed a theoretical framework incorporating global visual context for group activity recognition, consistent with transformer-based causal graphs. Li et al. [35] proposed a method using cross-connected transformers and clustered attention mechanisms to model spatial and temporal contexts. Kong et al. [36] further advanced this approach by predicting future group states based on long- and short-term historical state dynamics. These transformer-based models, utilizing supervised or semi-supervised learning, have consistently achieved state-of-the-art performance in GAR.

Traditional HAR research relies on manual annotations for supervised learning, but Nakanishi et al. [37] proposed a method for learning group activity features without such annotations. They treated activity classes and appearance features as person attributes, training a network in an end-to-end manner to predict group activity based on individual attributes. This approach effectively extracted multi-person activity features.

However, existing HAR research on operating room surveillance videos has primarily focused on human detection and pose estimation of surgical team members [38–41]. To date, no studies have investigated activity recognition for surgical teams. Therefore, it is necessary to explore HAR models specifically for intraoperative analysis.

1.5 Surgical Team Members Captured in Operating Room Surveillance Video

This section focuses on the specific activities and positioning of surgical team members captured in operating room surveillance videos. As shown in Figure 1.1 (a), surgical team members are positioned around the operating table according to their roles. Their primary responsibilities are as follows:

- Surgeon: Stands near the patient and performs the surgery, directly interacting with the surgical field.
- Anesthetist: Positioned at the patient's head, operates the anesthesia machine, and monitors the patient's condition.
- Scrub Nurse: Stands next to the surgeon, assisting by passing surgical instruments.
- Circulator: Moves within and outside the surgical field to deliver necessary medical instruments to the scrub nurse.

Next, we consider the activity features of surgical team members upon visually reviewing the operating surveillance video. Figure 1.1 (b) shows a sample frame from an actual operating room surveillance camera. The surgical team members' activity features are as follows:

- Often standing still.
- The positions are rarely changed except for the circulator.
- Often focusing on the surgical field during surgery.

Operating room surveillance videos present several challenges for activity recognition:

- Restricted camera angles limit the visibility of activities and interactions.
- The surgical field and operating table are occluded.
- Surgical team members are often obscured by medical instruments.
- Surgeries may last over 3 h, 30 min [42] making manual visual analysis impractical

Given these challenges, we focus on detecting large movements and postural changes to extract relevant information from surveillance video. While complete occlusion prevention is difficult, analyzing activities from a fixed camera angle can help develop simple intraoperative analysis techniques applicable to all operating rooms equipped with surveillance cameras.

1.6 Applications of Vision-Based HAR for Surgery

We consider the significant activities captured in operating room surveillance videos and their potential applications for improving the safety and efficiency of surgeries:

Safety Improvements Incident detection is essential for surgical safety improvement. Since the medical errors occur suddenly, we consider that predicting incidents becomes a challenging task. It is highly desirable to detect all events that may impact a patient's life, making recall an important evaluation metric. Therefore, pursuing real-time detection is necessary to identify incidents more quickly.

Surgical Education To enhance surgical training, it is essential for surgical specific activity recognition. For evaluation metrics for training, estimating the accuracy and speed of the activities is also essential. As educational materials, we extract surgical specific activities from stored operating room surveillance videos. Additionally, virtual reality (VR) and augmented reality (AR) technologies are necessary to perform simulations for training.

Surgical Evaluation In the evaluation of surgical procedures, it is essential to consider both real-time and post-operative metrics. Real-time evaluation metrics focus on assessing the appropriateness of the handles by each member or the team, based on their roles and the progress of the surgery. On the other hand, post-operative evaluation metrics involve qualitative assessments based on various criteria, including: the difference between the scheduled surgical time and the actual surgical time, the discrepancy between the estimated blood loss and the actual blood loss, and the presence or absence of post-operative complications.

To achieve these goals, two main research directions are required:

- **Recognition of Surgery-Specific Activities:** There are no publicly available datasets for intraoperative activity research. Since activities vary by role, dedicated activity models are needed for each surgical team member.
- **Intraoperative Situation Recognition:** If surgical procedure logs were synchronized with operating room surveillance videos, it would be possible to search and extract key scenes, such as the start, end, and emergency events, from large video datasets. Further analysis of these extracted datasets could drive research aimed at improving surgical safety and efficiency.

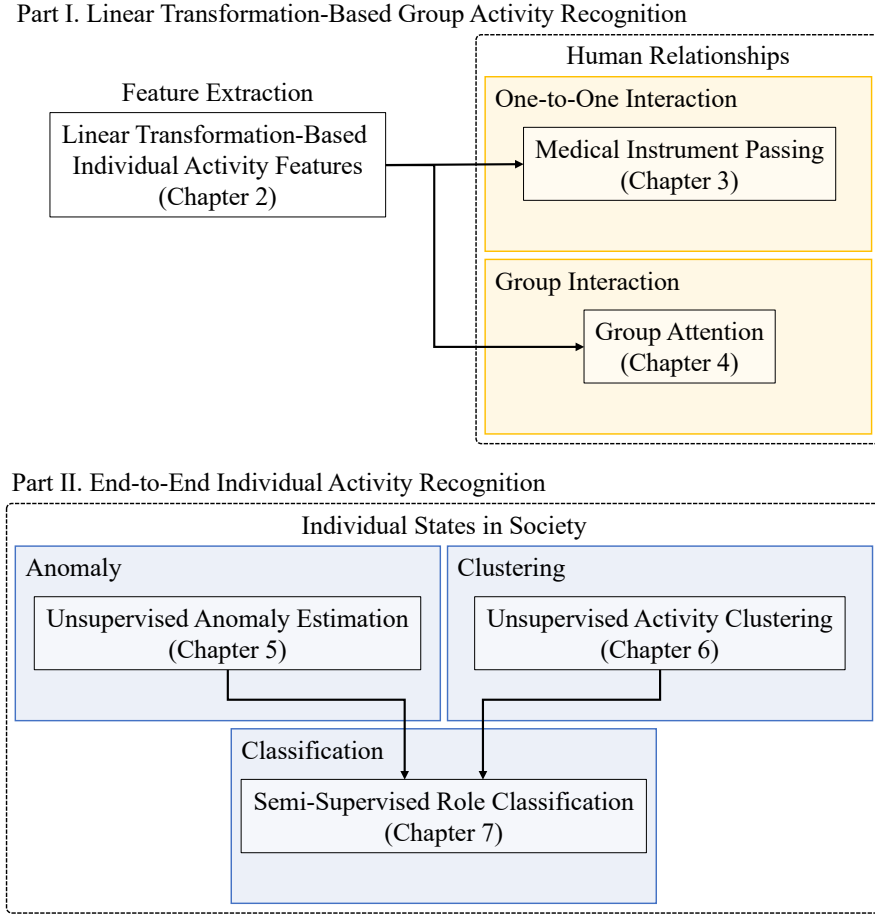


Figure 1.2: The research framework of vision-based HAR for operating room surveillance videos.

1.7 Research Scope and Conditions

This study focuses on recognizing surgical-specific activities using operating room surveillance video. Additionally, we apply our proposed activity recognition models to videos spanning the entire surgical process, from preparation to cleanup, to verify whether information such as changes in surgical state and the occurrence of emergencies can be extracted as logs.

Figure 1.2 shows the relationship and load-map of vision-based HAR conducted in this study. In Part I, we propose a method for recognizing interactions between surgical team members. Since our focus is on interaction recognition, we define individual activities as time series data of skeleton models obtained through pose estimation using a tracking model. We extract features of individual activities through linear transformation (Chapter 2). Two key group activities relevant to surgical procedures are analyzed: Medical Instrument Passing (Chapter 3) which involves the transfer of clean-packaged

surgical instruments between the circulating nurse and scrub nurse. Group Attention (Chapter 4) which may be drawn to unexpected events.

In Part II, we focus on extracting the states and features of individual activities. Although operating room surveillance videos are continuously recorded, identifying specific surgical activities within them is challenging. Most intraoperative activities involve standing still, leading to numerous ambiguous cases, making supervised learning annotation difficult.

We trained deep neural network models based on autoencoders and variational autoencoders (VAE) on a large volume of operating room surveillance video. Using the latent features extracted from these models, we analyze individual activities. First, we proposed the unsupervised anomaly estimation model of intraoperative activity to detect rare intraoperative activities (Chapter 5). Second, we proposed the unsupervised clustering model that groups intraoperative individual activities to analyze significant activity features (Chapter 6). Finally, we proposed the semi-supervised role classification model (Chapter 7), which recognizes role-specific activity patterns and how they change based on surgical events.

To ensure the generalizability of our proposed models across various operating rooms, we do not limit our analysis to specific surgical procedures. However, to minimize occlusions, we only use surveillance videos from operating rooms without large medical equipment such as CT machines and surgical robots. Additionally, since surveillance cameras vary in field of view due to room size and instrument arrangement, resulting in inconsistencies in camera angles and internal parameters, we do not perform camera calibration to reduce application costs.

1.8 Dissertation Overview

Figure 1.3 illustrates the dissertation structure, which consists of three main parts: (a) Linear Transformation-Based Group Activity Recognition, (b) End-to-End Individual Activity Recognition, and (c) Intraoperative Event Detection.

Part I focuses on group activity recognition using linear transformations. Chapter 2 details individual activity feature extraction through keypoint-based transformations. Chapter 3 examines medical instrument passing between the scrub nurse and circulator. Chapter 4 explores relationships among surgical staff without role recognition.

Part II addresses individual activity recognition using end-to-end models. Chapter 5 introduces an unsupervised anomaly estimation model that detects rare intraoperative

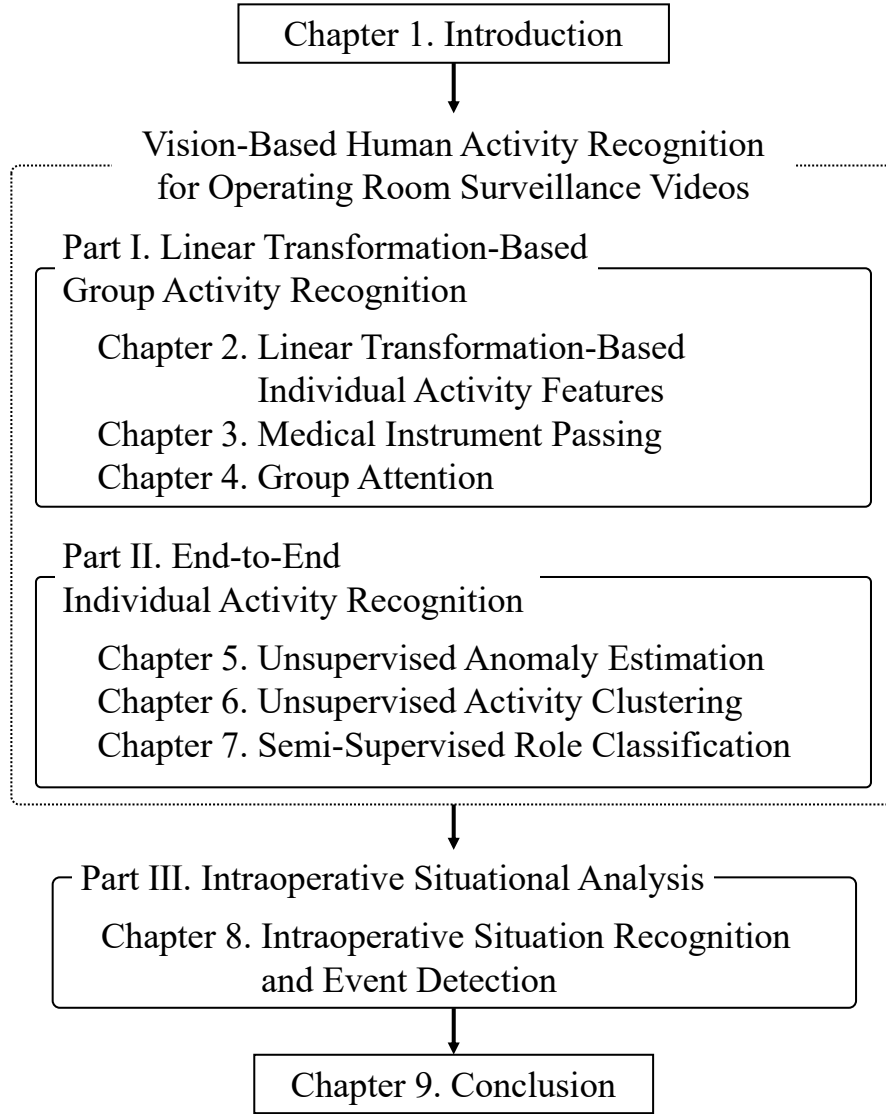


Figure 1.3: This dissertation comprises three main parts: a) Linear Transformation-Based Group Activity Recognition, b) End-to-End Individual Activity Recognition, and c) Intraoperative Analysis.

activities. Chapter 6 clusters activity features from autoencoders to infer roles. Chapter 7 integrates these methods for semi-supervised role classification.

Part Part III applies these methods to operating room surveillance videos, from preparation to cleanup, to analyze intraoperative situations. Section 8.2 evaluates the feasibility of detecting surgery phase transitions using group activity recognition. Section 8.3 explores event detection and situation recognition based on individual activity changes identified by the role classification model.

Finally, Chapter 9 summarizes the findings and discusses future research directions.

Part I

Linear Transformation-Based Group Activity Recognition

In this part, we describe the recognition of intraoperative interactions. Team coordination is a significant factor in the progression of surgery. To ensure smooth surgical procedures, the anesthetist manages the patient's vitals, the scrub nurse handles the necessary medical instruments, and the circulating nurse delivers additional instruments from outside the surgical field. Consequently, interactions occur among surgical team members. This part focuses on two types of interactions observable in operating room surveillance videos: a) Medical Instrument Passing, and b) Group Attention.

Medical Instrument Passing: The activity in which the circulator passes medical instruments to the scrub nurse can serve as an indicator of surgical progress. Since surgical procedures follow a predetermined sequence based on the type of operation, the progress of the surgery can be quantified by observing whether instruments are handed over according to the expected timeline. If an urgent event occurs, delays or deviations in instrument passing may be detected.

Group Attention: Detecting group attention provides insight into situational recognition, including whether the operation is proceeding smoothly or if an unexpected event has occurred. During routine surgery, surgical team members typically focus on different aspects: the surgeon and scrub nurse concentrate on the patient, the anesthetist monitors the anesthesia machine, and the circulating nurse oversees the entire operating room. If both the scrub nurse and circulating nurse turn their attention to the monitor, this may indicate an unexpected event requiring their focus.

To recognize these interactions, we extract features from the time series of keypoints corresponding to individual activities. These features can be obtained by processing operating room surveillance videos using object detection, pose estimation, and tracking models developed in previous studies. While deep neural network-based activity recognition models may vary in accuracy depending on the operating room environment, using simple linear transformations for feature extraction can help mitigate this variability.

Chapter 2 presents the feature extraction methods for individual activities using linear transformations. Chapter 3 introduces the recognition model for medical instrument passing from the circulator to the scrub nurse. Chapter 4 describes the estimation model for determining team members' focal points within the operating room based on their facial orientation.

Chapter 2

Linear Transformation-Based Individual Activity Features

2.1 Introduction

In this chapter, we present the feature extraction method for individual activities using linear transformation. We input the time series keypoints obtained by pose estimation model and tracking model.

In previous research, initial approaches to keypoints-based individual activity recognition using handcrafting features [43–46] did not consider the semantic connectivity of body joints. Recent approaches [47–52] have utilized spatial-temporal graphs to represent the spatial relationships of body joints. However, these models are trained using the supervised learning of specific datasets [44, 53], which was not applicable in our case because of the complexity of surgical situations in operating rooms.

On the other hand, we aim to recognize medical instrument passing and group attention. Medical instrument passing involves the circulator and scrub nurse facing each other, with their hand positions in close proximity. Group attention requires estimating the direction of gaze from face orientation. We considered effective individual activity features for recognizing these group activities.

First, we applied homography transformation to keypoints and projected them onto a top-down view of the operating room. This allows for the visualization of the distribution and changes in features within the operating room, providing features that are easy for humans to interpret. Next, from the key points in the top-down view, we

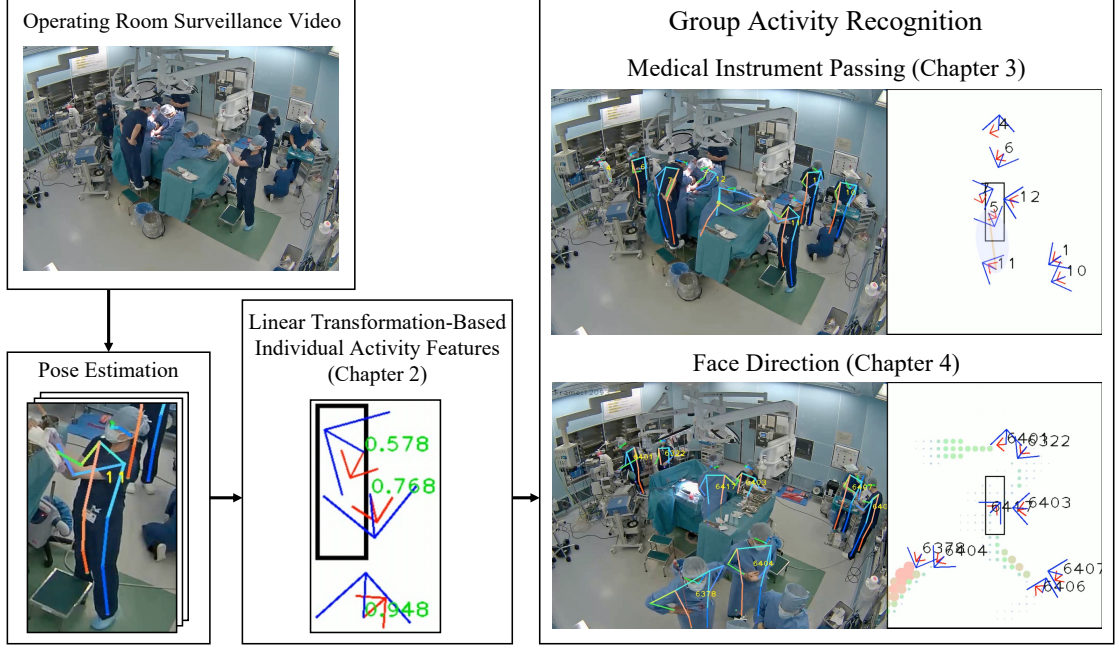


Figure 2.1: Data pipeline of group activity recognition comprising three main steps: 1) pose estimation, 2) individual activity feature extraction (Chapter 2), and 3) group activity recognition. We present two group activity recognition models: a) medical instrument passing (Chapter 3) and b) group attention (Chapter 4).

applied linear transformation to extract body position, body orientation, and face orientation. Additionally, we extracted arm flexion and wrist position from the operating room surveillance video (without homography transformation).

For medical instrument passing (Chapter 3), we used body position, body orientation, arm extension, and wrist position as input features. For group attention (Chapter 4), we used body position and face orientation as input features.

2.2 Data Pipeline of Group Activity Recognition

The scheme of the proposed data pipeline is shown in Figure 2.1. We used operating room surveillance videos as input and applied pose estimation to obtain individuals' keypoints. The format of the skeleton model used COCO keypoints [54], which include five facial points and twelve body points. We also used particle filters for tracking individuals to sort keypoints into time series data.

Applying the methods presented in this chapter, we extract individual positions, face orientations, body orientations, arm bending, and wrist positions. These features are then sent to recognize medical instrument passing and group attention as group activities.

2.3 Notations

Firstly, we describe about the notations for Part I (Chapter 2 - Chapter 4). We denote a 2D point by a bold uppercase letter such as $\mathbf{P}$. We use italic lowercase letters, such as d , to represent the scalar values. Specifically, we denote the abscissa and ordinate of a 2D point with italicized lowercase letters x and y , respectively. We denote a vector between two points using a bold lowercase letter with a right arrow, such as $\vec{\mathbf{v}}$.

An attribute is denoted by a subscript word, where _l and _r represent left and right, respectively. For example, $\mathbf{P}_{\text{ear_l}}(x_{\text{ear_l}}, y_{\text{ear_l}})$ denotes the left ear point and $\vec{\mathbf{v}}_{\text{face}}$ is a vector of the face orientation.

We denote a matrix with bold italic upper-case letters, such as $\mathbf{M}$. A set is denoted by a blackboard bold uppercase letter, such as $\mathbb{S}$.

$\max()$ denotes the maximum of several scalar values; for example, given $a > b$, $\max(a, b) = a$. $\text{mid}()$ denotes the midpoint of two points with left and right attributes, for example,

$$\text{mid}(\mathbf{P}_{\text{ear}}) = \frac{\mathbf{P}_{\text{ear_r}} + \mathbf{P}_{\text{ear_l}}}{2}.$$

The Euclidean distance between two points $\mathbf{P}, \mathbf{Q}$ is denoted by $\text{dis}(\mathbf{P}, \mathbf{Q})$. The cosine similarity of two vectors $\vec{\mathbf{v}}, \vec{\mathbf{u}}$ is denoted by $\cos(\vec{\mathbf{v}}, \vec{\mathbf{u}})$.

2.4 Pose Estimation

We applied pose estimation to operating room surveillance videos and extracted the keypoints that represent the skeleton of each individual.

In addition, we used the 17 points of COCO [54] for the keypoints, including 5 face points and 12 body points. Each point holds an estimation confidence score.

2.5 Tracking Using Particle Filter

Because pose estimation entails extracting and outputting the keypoints of individuals randomly, it is necessary to identify and continuously track each individual with consistency.

We tracked the midpoint $\text{mid}(\mathbf{P}_{\text{hip}})$ between the left and right hips for each individual because this afforded the best tracking stability for pose estimation. We applied a

particle filter with 300 particles to maintain individual tracking consistency. Each particle $\mathbf{X}_k$ followed a 2D Gaussian distribution $\mathbf{X}_k \sim N_2(\mu, \Sigma)$ where $\mu = \text{average}(\mathbf{X})$, $\Sigma = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix}$.

2.6 Individual Activity Features

We extracted the following six features for each individual: body position $\mathbf{P}_{\text{body}}$, face orientation $\vec{\mathbf{v}}_{\text{face}}$, body orientation $\vec{\mathbf{v}}_{\text{body}}$, arm flexion f_{arm} , left wrist position $\mathbf{P}_{\text{wrist.l}}$, and right wrist position $\mathbf{P}_{\text{wrist.r}}$.

2.6.1 Body Position

First, we calculated a homography matrix $\mathbf{H}$ based on the four fixed corners of the operating table to generate a 2D reprojection of the room. Given $\mathbf{H}$, we transformed the ankle points

$$\begin{bmatrix} \mathbf{P}'_{\text{ankle.r}} \\ \mathbf{P}'_{\text{ankle.l}} \end{bmatrix} = \mathbf{H} \begin{bmatrix} \mathbf{P}_{\text{ankle.r}} \\ \mathbf{P}_{\text{ankle.l}} \end{bmatrix}$$

and calculated the body position $\mathbf{P}_{\text{body}}$ as

$$\mathbf{P}_{\text{body}} = \text{mid}(\mathbf{P}_{\text{ankle}}). \quad (2.1)$$

If the confidence scores of $\mathbf{P}_{\text{ankle.r}}$ and $\mathbf{P}_{\text{ankle.l}}$ were less than 0.2, indicating that the estimation of ankle positions was not sufficiently reliable, we estimated the body position instead $\mathbf{P}_{\text{body}}$ using $\mathbf{P}_{\text{hip}}$ and $\mathbf{P}_{\text{shoulder}}$, as follows:

$$\vec{\mathbf{v}}_{\text{trunk}} = \text{mid}(\mathbf{P}_{\text{hip}}) - \text{mid}(\mathbf{P}_{\text{shoulder}}). \quad (2.2)$$

$$\mathbf{P}_{\text{body}} = \text{mid}(\mathbf{P}_{\text{hip}}) + \vec{\mathbf{v}}_{\text{trunk}} \cdot 1.2 \quad (2.3)$$

2.6.2 Face Orientation and Body Orientation

By solving $\mathbf{P}_{\text{body}} = \mathbf{A}_{\text{ear}} \cdot \text{mid}(\mathbf{P}_{\text{ear}})$ and $\mathbf{P}_{\text{body}} = \mathbf{A}_{\text{shoulder}} \cdot \text{mid}(\mathbf{P}_{\text{shoulder}})$, we calculated the matrices $\mathbf{A}_{\text{ear}}$ and $\mathbf{A}_{\text{shoulder}}$ to perform an affine transformation to map the ear and nose points to the shoulder points, respectively, of each individual, corresponding

to their body position as follows:

$$\begin{bmatrix} \mathbf{P}'_{\text{ear.l}} \\ \mathbf{P}'_{\text{ear.r}} \\ \mathbf{P}'_{\text{eye.l}} \\ \mathbf{P}'_{\text{eye.r}} \end{bmatrix} = A_{\text{ear}} H \begin{bmatrix} \mathbf{P}_{\text{ear.l}} \\ \mathbf{P}_{\text{ear.r}} \\ \mathbf{P}_{\text{eye.l}} \\ \mathbf{P}_{\text{eye.r}} \end{bmatrix}$$

and

$$\begin{bmatrix} \mathbf{P}'_{\text{shoulder.r}} \\ \mathbf{P}'_{\text{shoulder.l}} \end{bmatrix} = A_{\text{shoulder}} H \begin{bmatrix} \mathbf{P}_{\text{shoulder.r}} \\ \mathbf{P}_{\text{shoulder.l}} \end{bmatrix}.$$

Given a rotation matrix

$$\mathbf{R} = \begin{bmatrix} \cos \frac{\pi}{2} & -\sin \frac{\pi}{2} \\ \sin \frac{\pi}{2} & \cos \frac{\pi}{2} \end{bmatrix},$$

we calculated the face orientation $\vec{\mathbf{v}}_{\text{face}}$ of an individual as follows:

$$\vec{\mathbf{v}}_{\text{face}} = \begin{cases} \mathbf{R} \cdot (\mathbf{P}'_{\text{ear.r}} - \mathbf{P}'_{\text{ear.l}}), & \text{if } x_{\text{ear.l}} \leq x_{\text{nose}} \leq x_{\text{ear.r}} \\ \text{OR } x_{\text{ear.r}} \leq x_{\text{nose}} \leq x_{\text{ear.l}} \\ \text{mid}(\mathbf{P}'_{\text{eye}}) - \text{mid}(\mathbf{P}'_{\text{ear}}), & \text{else} \end{cases} \quad (2.4)$$

Body orientation $\vec{\mathbf{v}}_{\text{body}}$ was calculated as follows:

$$\vec{\mathbf{v}}_{\text{body}} = \mathbf{R} \cdot (\mathbf{P}'_{\text{shoulder.r}} - \mathbf{P}'_{\text{shoulder.l}}). \quad (2.5)$$

2.6.3 Arm Flexion

To obtain the arm flexion, we first calculated the vector of the body trunk $\vec{\mathbf{v}}_{\text{trunk}}$ using Eq. (2.2).

We then calculated the vectors of each arm as

$$\begin{bmatrix} \vec{\mathbf{v}}_{\text{arm.r}} \\ \vec{\mathbf{v}}_{\text{arm.l}} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{\text{wrist.l}} - \mathbf{P}_{\text{shoulder.l}} \\ \mathbf{P}_{\text{wrist.r}} - \mathbf{P}_{\text{shoulder.r}} \end{bmatrix}. \quad (2.6)$$

Given the body trunk and arm vectors, we estimated the likelihood of the flexion of each arm as follows:

$$\begin{bmatrix} f_{\text{arm.r}} \\ f_{\text{arm.l}} \end{bmatrix} = \begin{bmatrix} 1 - |\cos(\vec{\mathbf{v}}_{\text{trunk}}, \vec{\mathbf{v}}_{\text{arm.r}})| \\ 1 - |\cos(\vec{\mathbf{v}}_{\text{trunk}}, \vec{\mathbf{v}}_{\text{arm.l}})| \end{bmatrix}. \quad (2.7)$$

The arm flexion of an individual was determined by

$$f_{\text{arm}} = \max(f_{\text{arm.r}}, f_{\text{arm.l}}). \quad (2.8)$$

2.6.4 Wrist Position

We extracted the left and right wrist positions $\mathbf{P}_{\text{wrist.l}}$ and $\mathbf{P}_{\text{wrist.r}}$ from the keypoints of the pose estimation. These features were not transformed by any affine or homography transformation.

2.7 Summary of Individual Activity Features

These basic features of individuals

$$\mathbb{F}_{\text{individual}} = \{\mathbf{P}_{\text{body}}, \vec{\mathbf{v}}_{\text{face}}, \vec{\mathbf{v}}_{\text{body}}, f_{\text{arm}}, \mathbf{P}_{\text{wrist.l}}, \mathbf{P}_{\text{wrist.r}}\}$$

were sent to the estimation models to obtain the likelihood of each group activity in which the individual may be involved.

The extracted linear transformation-based individual features are shown in Figure 2.1; the blue arrows, the red arrows and the green values show the face orientations, body orientations and arm flexion, respectively.

In Chapter 3 and Chapter 4, we describe in details of medical instrument passing and group attention, respectively.

Chapter 3

Medical Instrument Passing

3.1 Introduction

In this chapter, we present the recognition model of medical instrument passing (MIP). The activity whereby the circulator passes medical instruments to the scrub nurse can be used as an indicator of the surgical procedure. As the surgical procedure is roughly determined by the type of operation, the progress of an operation can be quantified by observing whether medical instruments are handed over in accordance with the timing of the procedure. If an emergency occurs, there is a high possibility that the medical instrument passing will be detected as delayed or deviating from the expected procedure.

Using the individual activity features (body position, body orientation, arm flexion, wrist position) extracted in Chapter 2, we extracted features of the relative relationships between the individual activity features of all pairs of people in the operating room. Additionally, we trained the MIP recognition model with supervised learning.

3.2 Methods

3.2.1 Extracting Features

We estimated the activity of MIP based on the Euclidean distance of position, body orientation, average arm flexion and the Euclidean distance between the wrists of two individuals.

We first calculated the Euclidean distance of position $d_{p_a\&b}$ between the two individuals using

$$d_{p_a\&b} = \text{dis}(\mathbf{P}_{\text{body_a}}, \mathbf{P}_{\text{body_b}}), \quad (3.1)$$

and the minimum Euclidean distance of two individuals' wrists $d_{w_a\&b}$ using

$$d_{w_a\&b} = \min(\text{dis}(\mathbf{P}_{\text{wrist.l.a}}, \mathbf{P}_{\text{wrist.l.b}}), \text{dis}(\mathbf{P}_{\text{wrist.l.a}}, \mathbf{P}_{\text{wrist.r.b}}), \text{dis}(\mathbf{P}_{\text{wrist.r.a}}, \mathbf{P}_{\text{wrist.l.b}}), \text{dis}(\mathbf{P}_{\text{wrist.r.a}}, \mathbf{P}_{\text{wrist.r.b}})), \quad (3.2)$$

through which we obtained the minimum distance between all the pairs of wrists of two individuals.

The Euclidean distances of $d_{p_a\&b}$ and $d_{w_a\&b}$ follow 1-dimensional Gaussian distributions as $d_{p_a\&b} \sim N_1(\mu_p, \sigma_p)$ and $d_{w_a\&b} \sim N_1(\mu_w, \sigma_w)$, where $\mu_p = 200$, $\sigma_p = 100$, $\mu_w = 100$, and $\sigma_w = 80$, because the range of these values is $[0, 1]$. However, these parameters were set by manual tuning. This made it necessary to consider an optimization method.

We also evaluated the body orientation of two individuals in relation to each other $\mathbb{F}_a$ and $\mathbb{F}_b$ using

$$\begin{bmatrix} c_{ab} \\ c_{ba} \end{bmatrix} = \begin{bmatrix} \cos(\vec{\mathbf{v}}_{\text{body.a}}, \mathbf{P}_{\text{body.b}} - \mathbf{P}_{\text{body.a}}) \\ \cos(\vec{\mathbf{v}}_{\text{body.b}}, \mathbf{P}_{\text{body.a}} - \mathbf{P}_{\text{body.b}}) \end{bmatrix}. \quad (3.3)$$

Then, we calculated the likelihood of two individuals facing each other using

$$c_{a\&b} = \frac{c_{ab} + c_{ba}}{2}. \quad (3.4)$$

Because the range of $c_{a\&b}$ was $[-1, 1]$, we adjusted the range to $[0, 1]$ through min-max normalization as follows:

$$c_{a\&b} = \frac{c_{a\&b} + 1}{2}. \quad (3.5)$$

The average arm flexion was $f_{a\&b}$ obtained using

$$f_{a\&b} = \frac{f_a + f_b}{2}. \quad (3.6)$$

A feature set $\mathbb{S} = \{d_{p_a\&b}, d_{w_a\&b}, c_{a\&b}, f_{a\&b}\}$ was generated for each pair of individuals for each frame and was paired with a binary ground truth, which was manually input for training and validating the recognition model. The features of the MIP activity are shown in Figure 3.1.

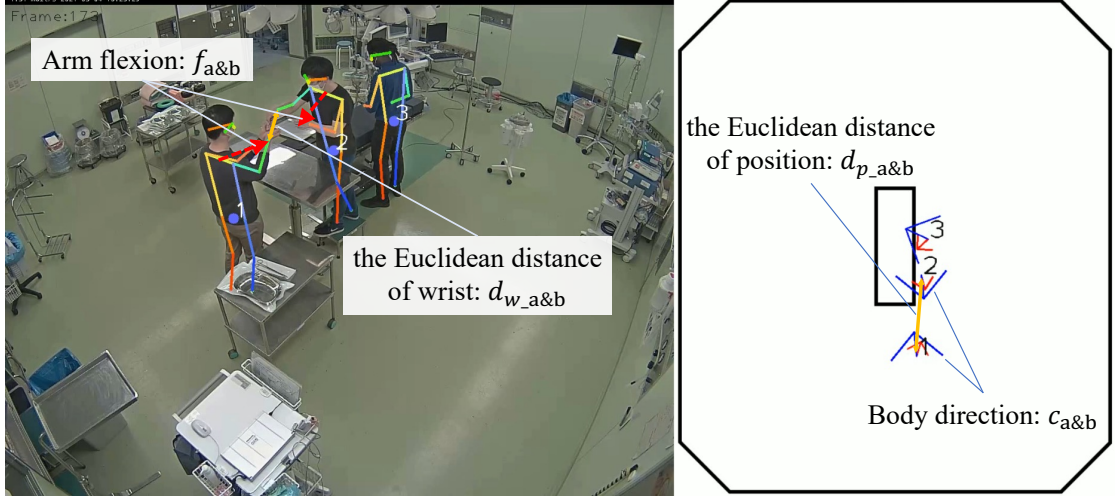


Figure 3.1: Features of MIP.

3.2.2 Machine-Learning-Based Models

We used two machine learning models to predict MIP activity. First, we used a support vector machine (SVM). We input a feature $\mathbb{S}$ into the SVM model, and the model classified the feature into one of two classes, based on whether the medical instrument was passed or not.

Second, we used an RNN using LSTM because the temporal changes of features can be candidate vital factors for activity prediction. Figure 3.2 shows the network architecture. We input features $\mathbb{S}$ of last T frames into the RNN model. The model yielded a 2D output that included likelihoods of passing and not passing medical instrument. We considered the max value of likelihoods as the prediction result of the model.

3.3 Experiments

In this section, the experiment conducted to evaluate the effectiveness of the models at predicting the MIP activities are described.

3.3.1 Simulated Videos

Because actual operating room surveillance videos is not suitable for the purpose of this experiment owing to complex on-site situations, we first used videos of actors simulating the target activities. All the videos were filmed in an operating room at the Kyoto University Hospital using an installed surveillance camera.

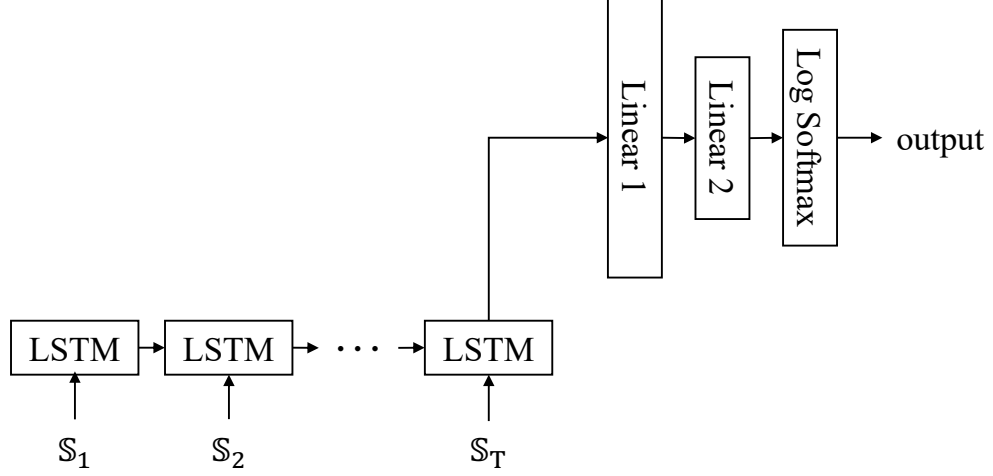


Figure 3.2: Architecture of RNN using LSTM. The model was composed of an LSTM layer, two linear layers, and a log-softmax layer. The features of the last T frames were input into the RNN model. The model yielded a 2D output that included the likelihoods of passing and not passing medical instrument.

3.3.1.1 Datasets

We recorded MIP activities from the circulating nurse to the scrub nurse, which is a regular interaction between dirty and clean areas during surgery. The actors were guided by experienced nurses to ensure that on-site concerns were accounted for.

Three actors played the roles of a surgeon, a scrub nurse, and a circulating nurse. The surgeon, who was not involved in MIP, was located in the clean area performing the surgery. The scrub nurse was also located in the clean area and received medical instruments from the circulating nurse. The circulating nurse was located in the dirty area and passed on sterilized and clean-packaged medical instruments to the scrub nurse. Both the scrub nurse and circulating nurse made their best efforts not to contaminate the instruments. First, the circulating nurse held the package up in front of them and ensured that no part of their bodies touched it. Then, the scrub nurse reached into the sterilized pack without touching its surface and extracted only the instrument.

The process was conducted using three different-sized instruments, syringe (small), scissors (medium), and tray (large). Regardless of size, passing each instrument engenders unique concerns that are reflected in the passing activity. We also recorded two different positions: **IP1** (orange) and **IP2** (violet), as shown in Figure 3.3. **IP1** was iterated twice and **IP2** was only once. The first iteration of **IP1** was used as the training data to train the SVM model and the other iterations were used as the test data. The composition of datasets by frame number of MIP activities are shown in Table 3.1.

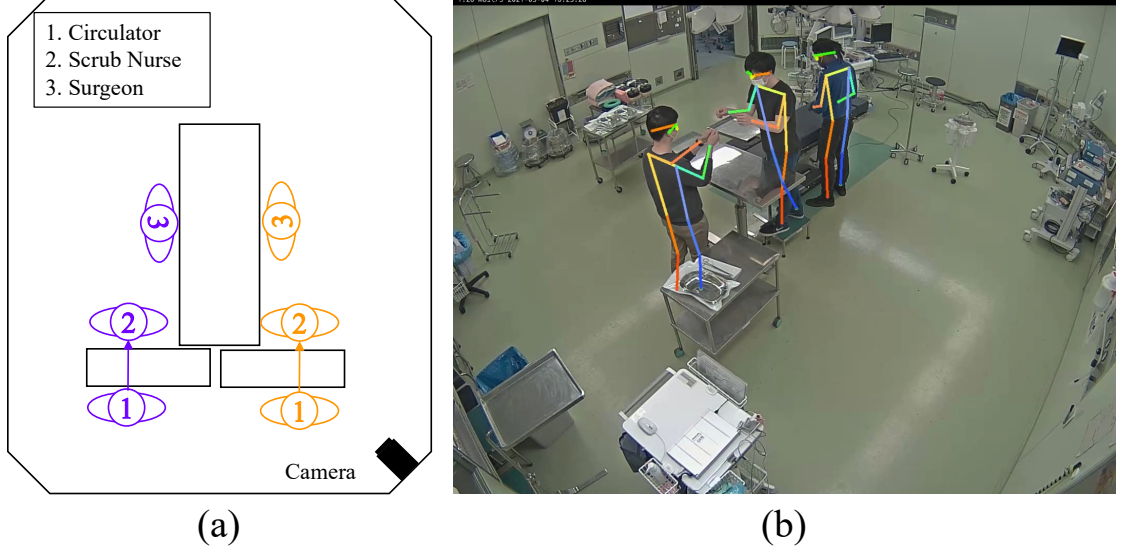


Figure 3.3: Images showing (a) the positions for MIP **IP1** (orange) and **IP2** (violet), and (b) the actual video frame of **IP1**.

3.3.1.2 Implementation

We used an SVM model to predict MIP in the simulation experiments. We utilized a radial basis function kernel for the SVM, with the gamma parameter γ and cost parameter C set at 0.1 and 0.7, respectively. These parameters were determined by the grid search algorithm, which was selected by γ : [0.01, 1.0] per 0.01, and C : (0.1, 10) per 0.1. However, we did not use the RNN model because the simulation data was too short for learning the RNN model.

3.3.1.3 Results

Figure 3.4 shows the results of the SVM model for predicting MIP activities in comparison with the ground truth. The solid blue and dashed orange lines represent the prediction result and ground truth, respectively. The other lines represent other features, in terms of “distance”: the Euclidean distance of position between the two individuals; “body”: the body orientation of two individuals in relation to each other; “arm”: the

Table 3.1: Composition of datasets by frame number of MIP activities.

Dataset	Frame Number of MIP	Frame Number of Not MIP	Total
training	538	1020	1558
IP1	447	4908	5355
IP2	715	5351	6066
Total	1700	11279	12979

average arm flexion; and “wrist”: the minimum Euclidean distance of wrists between the two individuals.

For position **IP1** (top), the model accurately predicted a narrow range as the ground truth, regardless of the size of the instrument. For position **IP2** (bottom), the prediction was the same as the ground truth for small and medium instruments and a wider range than the ground truth for large instruments.

The prediction results are listed in detail in Table 3.2. For the **IP1** dataset, the model predicted the activity with a recall score of approximately 45%. However, the predictions were within the ground truth regardless of the size of the medical instruments, as indicated by the top of Figure 3.4. For the dataset of **IP2**, the recall score of the model’s prediction of the activity was greater than 95%, which demonstrated the validity of the selected features with the SVM model for recognizing instrument passing activities from the simulated videos.

3.3.2 Real Operating Room Surveillance Videos

3.3.2.1 Datasets

We used three videos recorded in three operating rooms, denoted as **S1**, **S2**, and **S3**, respectively. We visually extracted the circulating nurse passing medical instruments to the scrub nurse from three actual operating room surveillance videos. The activity is a regular interaction between the dirty and clean areas during surgery.

We cropped the visually recognizable MIP activities from three operating room surveillance videos. Table 3.3 shows summary of the cropped data. The number of MIPs is the number of times MIP activities occurred in each surgery.

We split the data into training and test data in the proportion of 7 : 3 through random sampling (random seed: 64). Table 3.4 shows the composition of datasets by number of MIP activities. The training data contained 50 MIP activities and 5583 Non-MIP activities. The test data contained 18 MIP activities and 2375 Non-MIP activities. The ratio of the total training data to the total test data was approximately 7 : 3. Both

Table 3.2: Prediction results of the MIP activity.

Disposition	Accuracy [%]	Precision [%]	Recall [%]	F- measure
IP1	93.9	71.1	45.2	0.55
IP2	95.1	71.9	96.1	0.82
Average	94.5	71.5	70.7	0.69

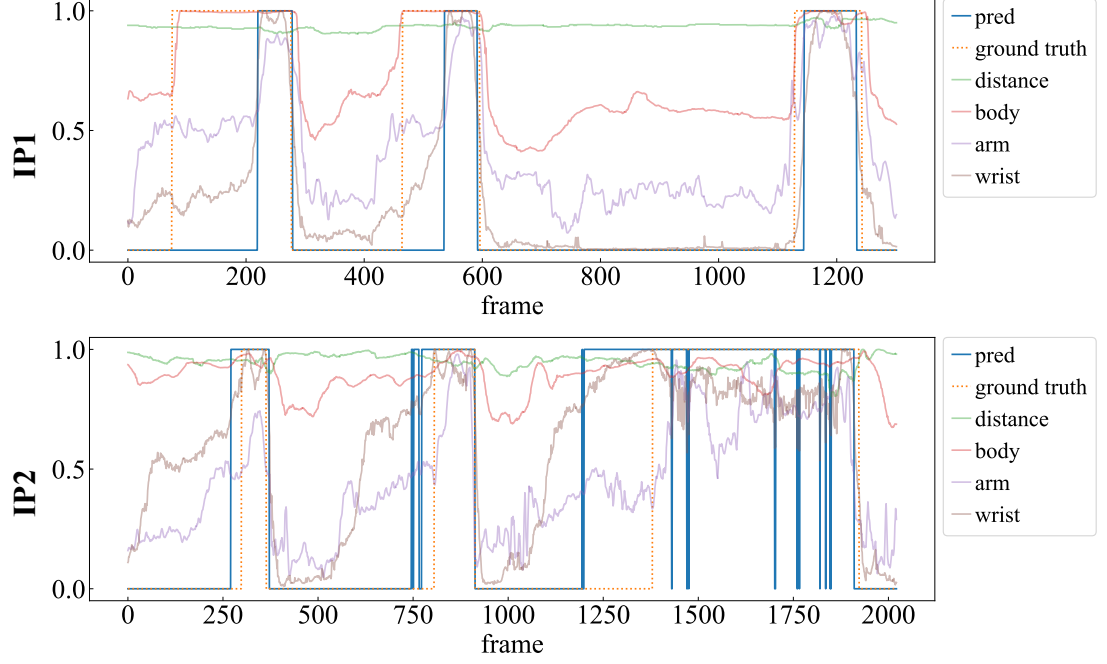


Figure 3.4: Results of the SVM model for predicting instrument passing activities in comparison with the ground truth. The solid blue line represents the prediction result and the dashed orange line represents the ground truth. Other lines represent other features in terms of “distance”: the Euclidean distance of position between the two individuals; “body”: the body orientation two individuals in relation to each other; “arm”: the average arm flexion; and “wrist”: the minimum Euclidean distance between the wrists of two individuals.

the training and test data had a significant bias of Non-MIP activity because there were many combinations of STs who did not perform MIP.

3.3.2.2 Implementation

We used an SVM model to predict MIP activities in simulation experiments. We utilized a radial basis function kernel for the SVM with gamma parameter γ and cost parameter C set at 0.02 and 0.1, respectively. We also adjusted the ratio of the MIP and Non-MIP activities training datasets in the training dataset such that it was 1 : 3 to prevent overfitting of the SVM model. When we adjusted the training datasets, we split the

Table 3.3: Summary of MIP data. The number of MIPs is the number of times MIP activities performed in each surgery.

Disposition	Number of MIP
S1	20
S2	39
S3	9
Total	68

datasets per frame, extracted all the MIP data and randomly sampled the Non-MIP data (64 random seeds). The number of frames of training datasets for learning the SVM model is shown in Table 3.5. These parameters were determined by a grid search algorithm that was selected by γ : [0.01, 1.0] per 0.01, C : (0.1, 10) per 0.1, and the ratio of MIP to Non-MIP activities: (1, 10) per 1.

In addition, we used an RNN model with LSTM, as shown in Figure 3.2. We set the number of hidden layers of the LSTM at 512, the number of the hidden layers of Linear 1 at 128, and the number of the hidden layers of Linear 2 at 8. The loss function for the learning model was binary cross entropy with logistic loss, which was numerically stable for the learning model because of the log-sum-exp trick. The loss function aimed to provide positive examples for learning biased data equally. We set the weight at (1, 16). The optimizer for the learning model was Adam [55] with learning rate $lr = 0.001$ and betas $(\beta_1, \beta_2) = (0.9, 0.999)$. We used all the training data for the learning model, which was different from when learning the SVM model.

3.3.2.3 Results

The prediction results of both models per frame for the test datasets are shown in Table 3.6. The recall scores per frame of both the SVM and RNN were 36.6% and 53.5%, respectively; therefore, the RNN predicted less loss than the SVM did.

However, the evaluation per frame was too strict to evaluate the timing of the MIP because the videos were 30 FPS. Therefore, we conducted the evaluation per event

Table 3.4: Composition of datasets by number of MIP activities. The training data had 50 MIP activities and 5583 Non-MIP activities. The test data included 18 MIP activities and 2393 Non-MIP activities. The ratio of the total training data to the total test data was approximately 7 : 3. Both the training data and test data contained significantly more Non-MIP activity because there were many combinations of STs not performing MIP.

Dataset	Number of MIP	Non-MIP activities	Total
Training	50	5583	5633
Test	18	2375	2393
Total	68	7958	8026

Table 3.5: Composition of training datasets by number of frame of MIP activities in actual surgical video for learning SVM model.

Datasets	Number of MIP	Number of Non-MIP	Total
Training	1554	4662	6216

whether there were MIP activities or not. We considered the beginning to the end of a video for each combination of individuals as a single MIP event.

Table 3.7 shows the prediction results of both models per event for the test datasets. The recall score per event of the RNN model was over 94%, indicating that the RNN model accurately predicted most of the MIP activities. However, the precision score per event of the RNN model was low, as was the evaluation per frame. Nevertheless, the ratio of the true negatives to the false negatives was $2332 : 43 \sim 54 : 1$, indicating that the errors in detecting passing were too small, in relation to the length of the entire test dataset.

Figure 3.5 shows the sample results of one case of MIP activity that the SVM model, unlike the RNN model, could not predict. The prediction made by the RNN model was almost the same as the ground truth; the SVM model, on the other hand, failed to predict this throughout the video. Therefore, adding temporal factors to the features for predicting MIP activities was considered effective.

Figure 3.6 shows the samples of actual frames of operating room surveillance videos. The MIP activities in both videos were accurately extracted. For the top, the MIP activities between the circulator (Individual 1) and scrub nurse (individual 3) and those between the circulator (Individual 2) and scrub nurse (Individual 4) were marked. We confirmed that the RNN model effectively extracted the MIP activities from actual operating room surveillance videos.

Table 3.6: Prediction scores of the MIP activity of test datasets per frame.

Disposition	Accuracy [%]	Precision [%]	Recall [%]	F- measure
SVM	99.5	35.5	36.6	0.36
RNN	98.8	22.7	53.5	0.32
Average	99.1	29.1	45.1	0.34

Table 3.7: Prediction scores of the MIP activity of test dataset per event.

Disposition	Accuracy [%]	Precision [%]	Recall [%]	F- measure
SVM	98.4	27.7	72.2	0.40
RNN	95.2	28.3	94.4	0.44
Average	96.8	28.0	83.3	0.42

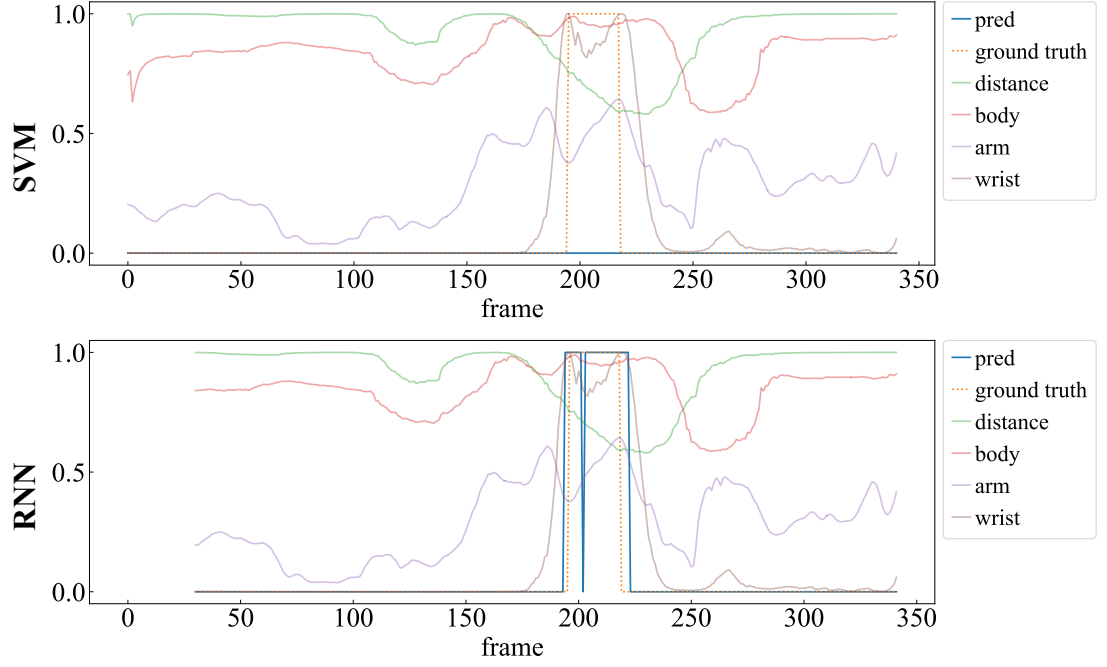


Figure 3.5: Sample results of one case of MIP activities the SVM model could not predict while RNN model could. The solid blue line represents the prediction result and the dashed orange line represents the ground truth. Other lines represent features in terms of “distance”: the Euclidean distance of position between the two individuals; “body”: the body orientation between two individuals; “arm”: the average arm flexion; and “wrist”: the minimum Euclidean distance between the wrists of two individuals.

3.4 Discussion

We used the SVM model for experiments using simulation and actual operating room surveillance videos. The performance of the model on both videos varied. For instance, the average recall scores of the SVM model on the simulation and actual videos were 70.7% and 36.6%, respectively. Actual operating room surveillance videos have more variety of patterns of placements than simulated cases. Figure 3.7 shows the scatter plots of features $\mathbb{S}$. In the simulated videos, the feature combinations of body distance and arm flexion, body orientation and arm flexion, and arm flexion and wrist distance of MIP and Non-MIP activities were clearly distinct. However, in the actual operating room surveillance videos, the features were not different. Therefore, the separating the hyperplane of the SVM proved inadequate.

Through experiment on an actual surgical video, we established that the RNN effectively learned about temporal changes in the features. We input the features of the last of 30 frames into the model. The recall score per frame improved by 16.9% and the recall score per event increased by 22.2%. Therefore, using the temporal context of features effectively improved the prediction results.

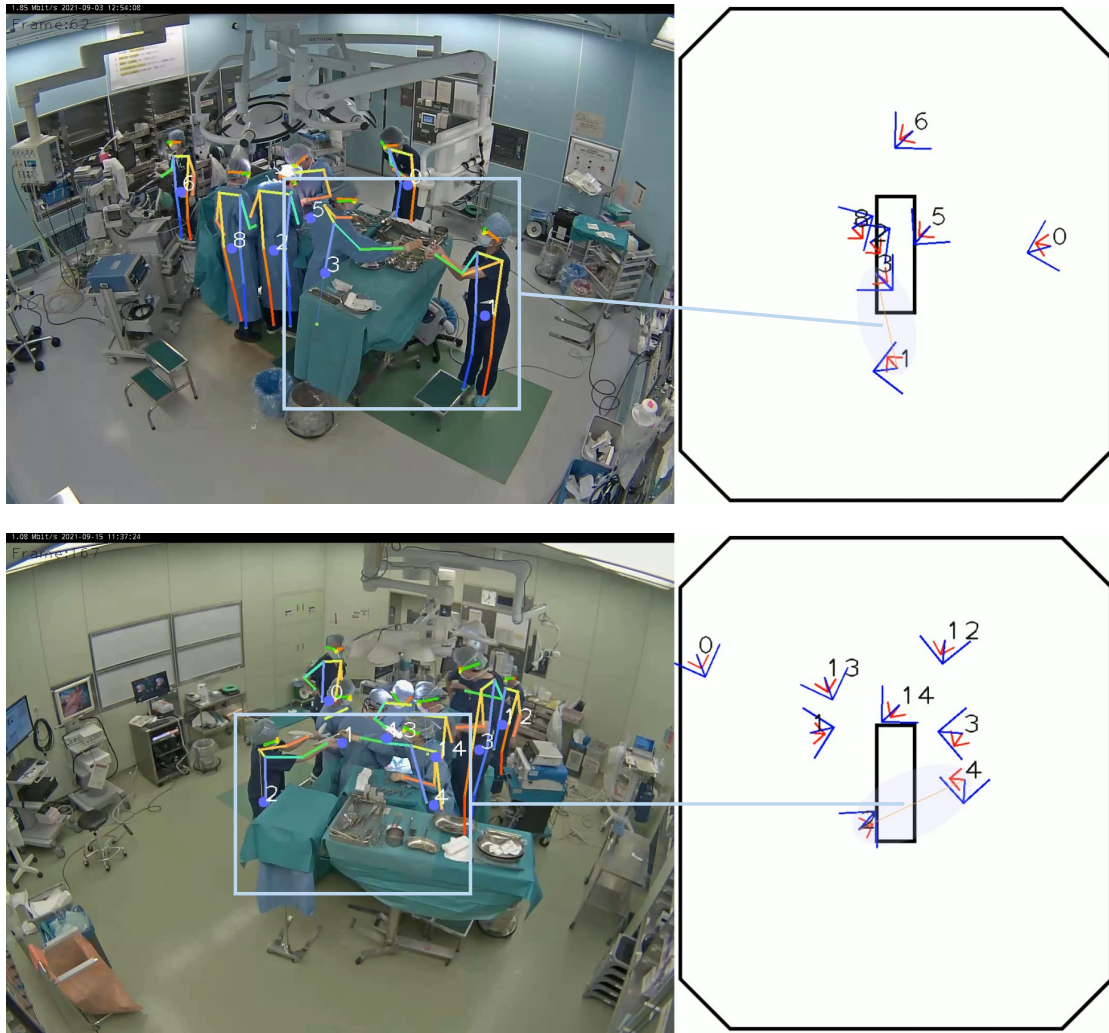


Figure 3.6: Samples of actual frames of operating room surveillance videos. For the top, MIP was marked between the circulator (Individual 1) and the scrub nurse (Individual 3). At the bottom, MIP was marked between the circulator (Individual 2) and scrub nurse (individual 4).

However, when an individual was occluded by equipment, the RNN model could not predict the passing activity correctly, as shown in Figure 3.8. Therefore, another model has to be used to predict the activity of occluded members based on past activity features; another option is to install another camera in the operating room to cover the occlusion area. Additionally, it is difficult to recognize objects in the operating room surveillance camera footage as it does not clearly show which medical instruments are being handed over. We believe this issue can be resolved by making changes such as using high-resolution cameras capable of recording 4K video or installing cameras above the scrub nurse.

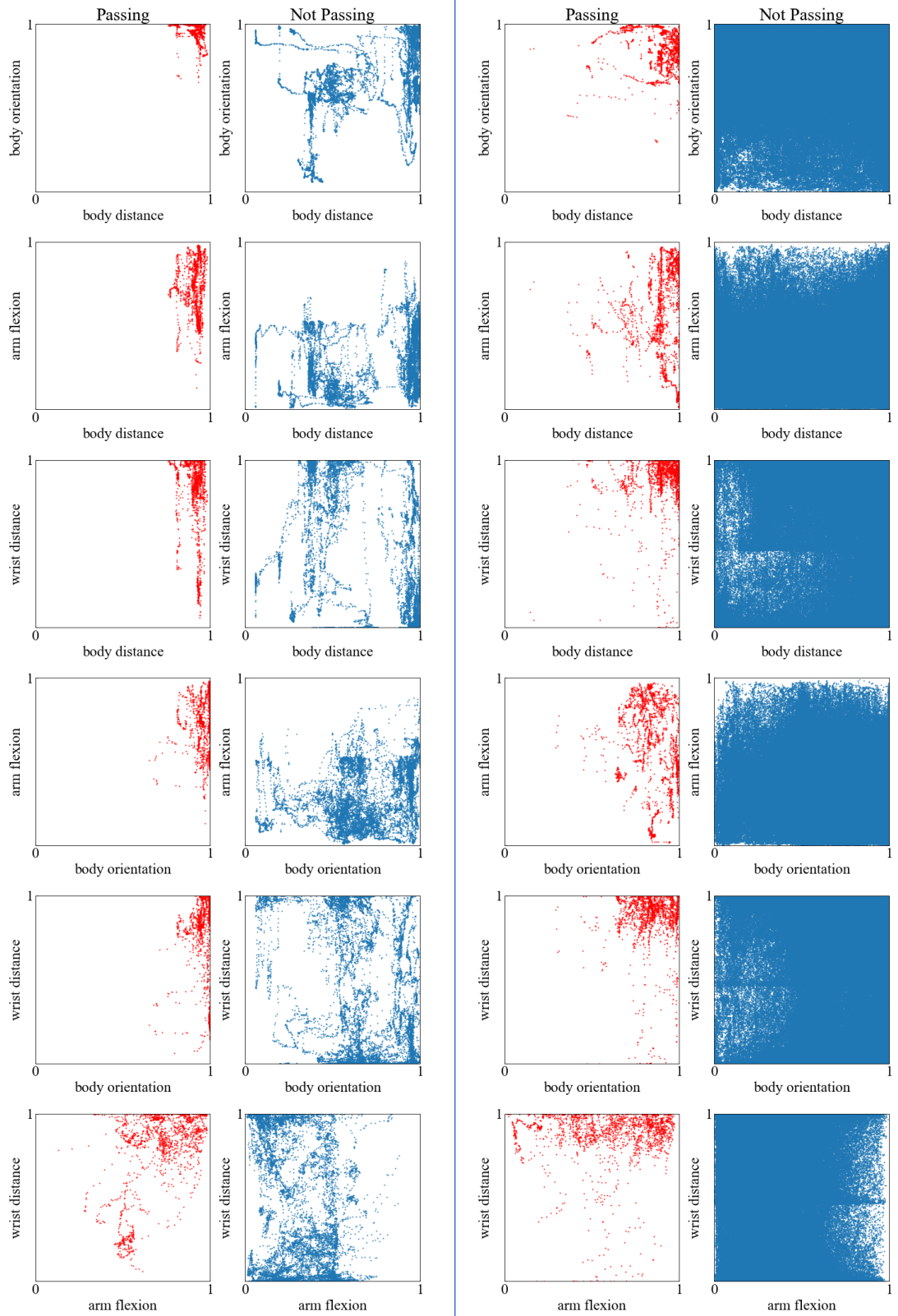


Figure 3.7: Scatter plots of MIP features $\mathcal{S}$, Left: simulated videos; Right: actual operating room surveillance videos. Red and blue points are MIP activity features and Non-MIP activity features, respectively.

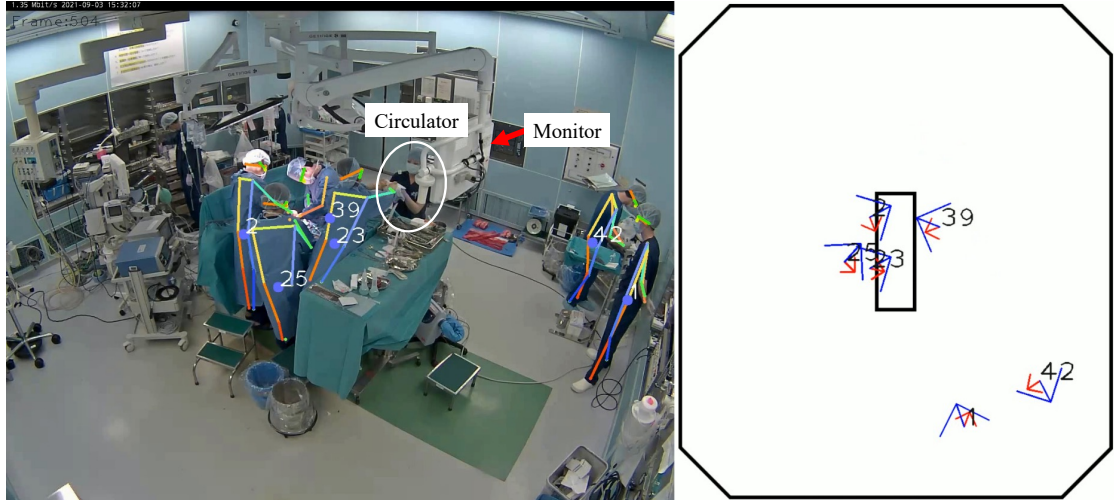


Figure 3.8: Sample failure result of estimation of operating room surveillance videos. The circulator was occluded by the monitor and was, therefore, undetected by pose estimation.

3.5 Conclusion

In this chapter, we proposed an MIP recognition model using individual activity features extracted in Chapter 2. In the experiments, we prepared and trained the model with both a simulation dataset and a dataset of actual surgical videos. The model was able to recognize MIP with high accuracy for both datasets. Additionally, using the proposed method, we found that it is also possible to recognize collaborative activities performed by two people facing each other.

Chapter 4

Group Attention

4.1 Introduction

In this chapter, we present the group attention (GA) estimation methods. During the smooth progress of surgery, attention is mostly focused on the surgical field. However, during preparation, cleanup, or emergencies, when many surgical team members are entering and exiting, attention is likely directed away from the surgical field. We considered that by estimating the gaze direction from surgical team members' facial direction and detecting the objects or areas being focused on, it is possible to detect events occurring during surgery.

We use the individual activity features (body position, face orientation) extracted in Chapter 2. From the time series positions and face orientations of each individual, we estimate where their gaze is directed on the top-down view of the operating room. Using the gaze estimation results for all individuals in the operating room, we visualize the heat-map of focused area on the top-down view of the operating room. We visualize how much attention the coordinates of surgical-related objects manually provided are received.

4.2 Methods

Given a point $\mathbf{X}$, the viewing length $l_{\mathbf{X}}$ of individual $\mathbf{I}_i$ and the viewing angle $\theta_{\mathbf{X}}$ of individual $\mathbf{I}_i$ are defined as follows:

$$l_{i|\mathbf{X}} = \text{dis}(\mathbf{X}, \mathbf{P}_{\text{body-}i})$$

$$\theta_{i|\mathbf{X}} = \cos(\vec{\mathbf{v}}_{\text{face-}i}, \mathbf{X} - \mathbf{P}_{\text{body-}i})$$

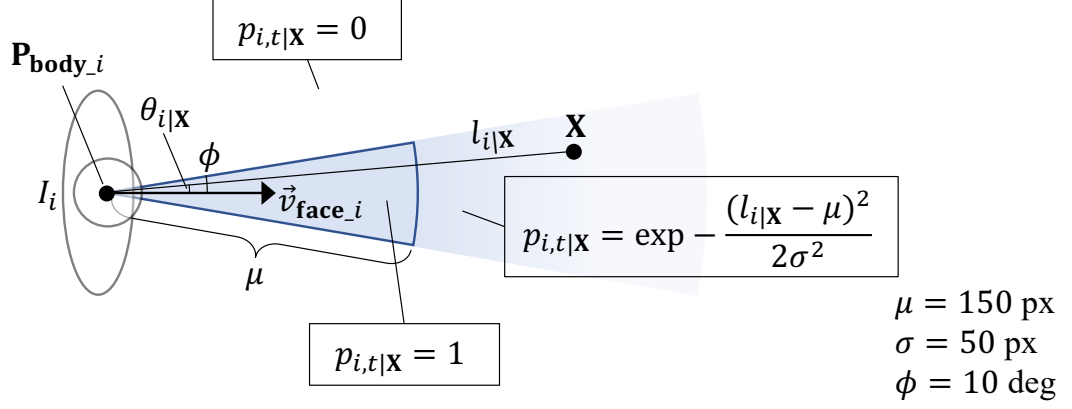


Figure 4.1: Diagram of viewing field of individual $\mathbf{I}_i$ at Frame t . A point $\mathbf{X}$ in the example in this figure, $|\theta_{i|\mathbf{X}}| \leq \phi$, and $l_{i|\mathbf{X}} > \mu$, where $\phi = 10 \text{ deg}$ and $\mu = 150 \text{ px}$. Thus, $p_{i,t|\mathbf{X}}$ was calculated using the middle case of Equation 4.1.

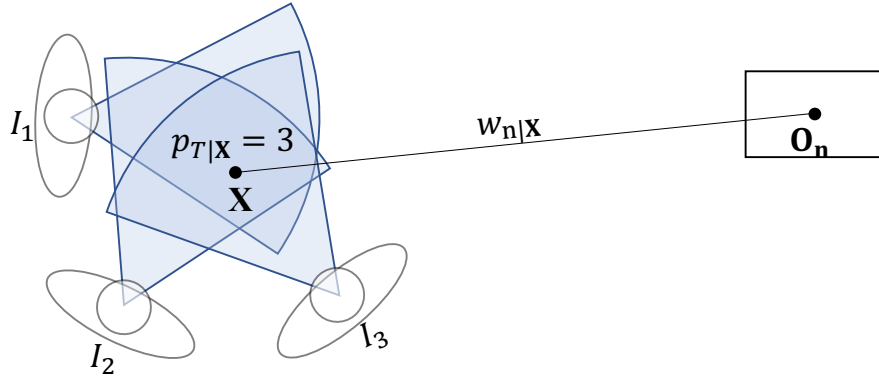


Figure 4.2: Diagram showing estimating attention score $s_{n|T}$ of $\mathbf{O}_n$ at frame T . In the example shown in this figure, there were three individuals, $\mathbf{I}_1$, $\mathbf{I}_2$, and $\mathbf{I}_3$. The attention probability of $p_{T|\mathbf{X}} = 3$ because $\mathbf{X}$ was in the view field of all three individuals.

where i represents the individual number.

The attention probability $p_{i,t|\mathbf{X}}$ of a point $\mathbf{X}$ by individual $\mathbf{I}_i$ at frame t was

$$p_{i,t|\mathbf{X}} = \begin{cases} 1, & \text{if } |\theta_{i|\mathbf{X}}| \leq \phi \text{ and } l_{i|\mathbf{X}} \leq \mu \\ \exp - \frac{(l_{i|\mathbf{X}} - \mu)^2}{2\sigma^2}, & \text{else if } |\theta_{i|\mathbf{X}}| \leq \phi \text{ and } l_{i|\mathbf{X}} > \mu \\ 0, & \text{else} \end{cases} \quad (4.1)$$

where $\mu = 150 \text{ px}$, $\sigma = 50 \text{ px}$, and $\phi = 10 \text{ deg}$. Figure 4.1 shows the diagram of viewing field.

The attention probability $p_{T|\mathbf{X}}$ of a point $\mathbf{X}$ at frame T was aggregated as the moving average of the attention probability of all the individuals in the last 30 frames (1 sec).

$$p_{T|\mathbf{X}} = \frac{1}{30} \sum_{t=T-30}^T \sum_{i \in I} p_{i,t|\mathbf{X}} \quad (4.2)$$

In addition, we assigned the object points that were recognized visually in actual operating room surveillance videos. We estimated how individuals paid attention to an object in the operating room.

$w_{n|\mathbf{X}}$ gives a weight between an object point $\mathbf{O}_n = (x_o, y_o)$ and a point $\mathbf{X}$ by

$$w_{n|\mathbf{X}} = \exp - \frac{(x - x_o)^2 + (y - y_o)^2}{2\sigma^2} \quad (4.3)$$

where $\sigma = 50$ px.

Thus, the attention score s_n of an object point $\mathbf{O}_n$ at frame T was calculated based on the weighted average as follows:

$$s_{n|T} = \frac{\sum_{\mathbf{X} \in \mathbb{R}^2} w_{n|\mathbf{X}} \cdot p_{T|\mathbf{X}}}{\sum_{\mathbf{X} \in \mathbb{R}^2} w_{n|\mathbf{X}}}, \quad (4.4)$$

Figure 4.2 shows the diagram of the estimated attention score $s_{n|T}$.

4.3 Experiments

The effectiveness of the proposed models for predicting GA activities was investigated in experiment on group activity, which is described in this section.

4.3.1 Simulated Videos

4.3.1.1 Datasets

Because real operating room surveillance videos is not suitable for the purpose of this experiment due to complex on-site situations, we first use videos simulated by actors conducting the target activities. For group attention, we recorded videos in which multiple actors gazed at different objects and on the same object. A monitor $\mathbf{O}_1$ and a desktop computer $\mathbf{O}_2$ were placed at the two corners of the room. An operating table $\mathbf{O}_3$ was positioned in the center of the room, as shown in Figure 4.3.

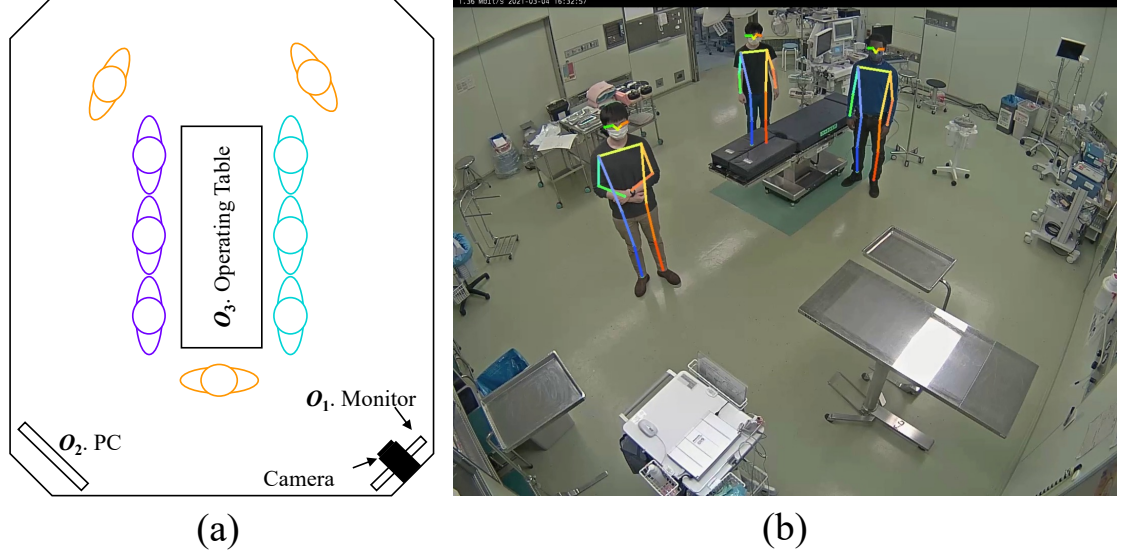


Figure 4.3: Images (a) showing the positions for group attention **GA1** (orange), **GA2** (violet), and **GA3** (light blue), and (b) the actual frame of **GA1**.

Each iteration took 20s and in the first 10s, all three actors were looking at different objects O_1 , O_2 , and O_3 . However, in the last 10s, they focused on the same objects, O_1 , O_2 , and O_3 , according to the iteration. We also instructed the actors to perform iterations with three placements, **GA1** (orange), **GA2** (violet), and **GA3** (light blue), as shown in Figure 4.3, which resulted in $3 \times 3 = 9$ datasets in total. As there was no training process for the group attention estimation, all the data were sent to our estimation model as test data.

4.3.1.2 Results

Figure 4.4 shows the histograms of the attention score results. The top, middle, and last row show the results for **GA1**, **GA2**, and **GA3**, respectively. The legend, “ O_n -N/A”, denotes that the actors were **not** paying **attention** to O_n and “ O_n -A” paying **attention** to O_n .

Across all the experiments, the mean of the distribution of the attention scores of O_3 was higher than that of O_1 and O_2 . Specifically, the mean of both O_1 -A and O_1 -N/A was approximately 0.1, the mean of O_2 -A was approximately 0.1, and that of O_2 -N/A was approximately 0.5; the mean of both O_3 -A and O_3 -N/A is approximately 1.0.

We confirmed that there was a significant difference between O_n -A and O_n -N/A in each result to clearly determine whether attention is being paid to an object or not. We set the null and alternate hypotheses as $\mu_{n-A} = \mu_{n-N/A}$ and $\mu_{n-A} > \mu_{n-N/A}$, respectively, and tested the one-sided Mann-Whitney U test. Table 4.1 shows the

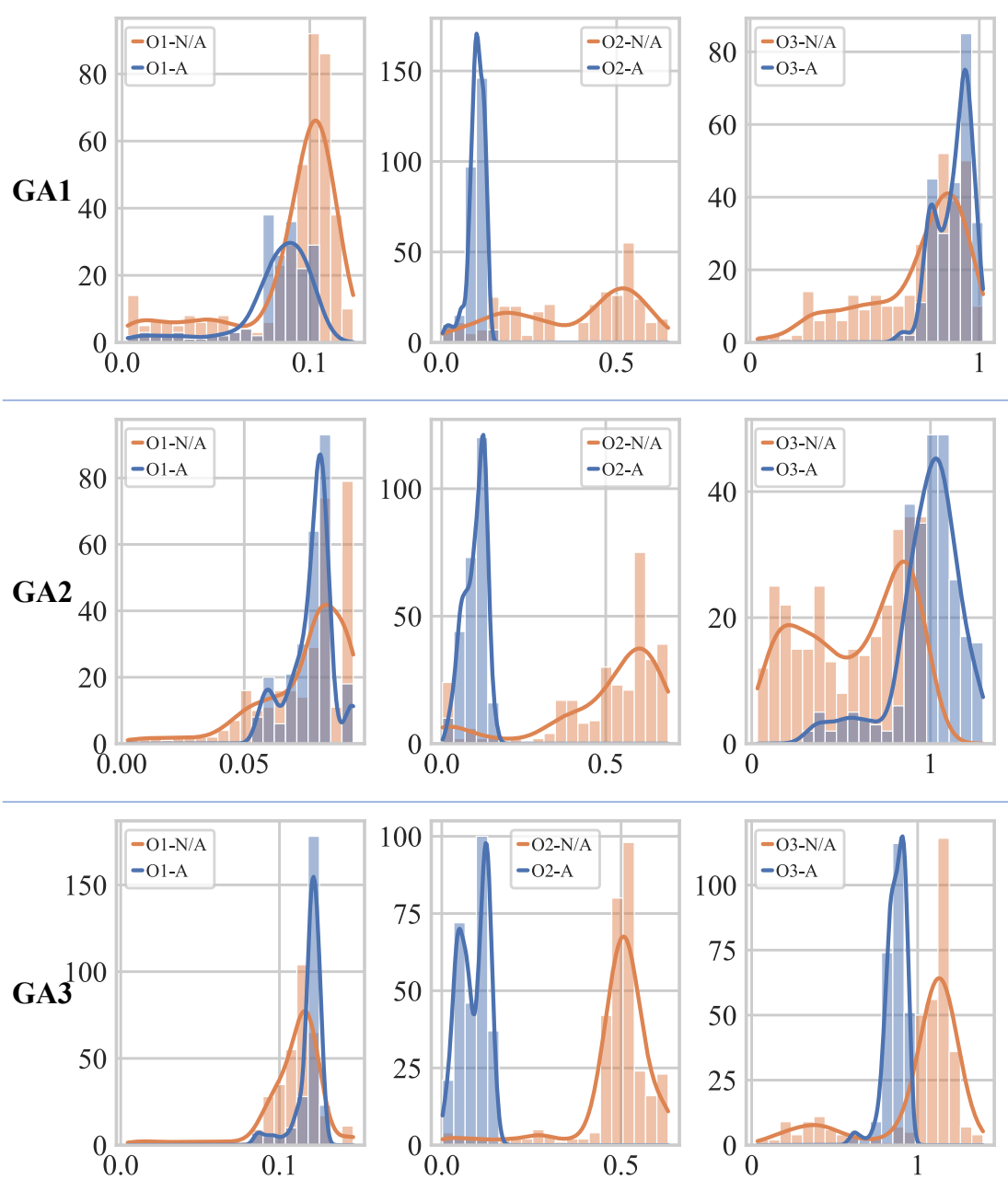


Figure 4.4: Histograms of attention score results. The top, middle, and last rows show the results for **GA1**, **GA2**, and **GA3**, respectively. The legend of “ O_n -N/A” denotes that the actors were **not** paying **attention** to O_n and “ O_n -A” paying **attention** to O_n , where $n = 1, 2, 3$.

results of the p-values obtained through the Mann-Whitney U test. The p-values of $\mathbf{O}_3$ for **GA1**, $\mathbf{O}_3$ for **GA2**, and $\mathbf{O}_1$ for **GA3** were $p < 0.05$; therefore, the null hypothesis was rejected. There were significant differences in $\mu_{n-A} > \mu_{n-N/A}$ for the attention scores between $\mathbf{O}_{n-A}$ and $\mathbf{O}_{n-N/A}$. However, because the p-values of the others were $p > 0.05$, the null hypothesis was accepted and there were no significant differences to $\mu_{n-A} > \mu_{n-N/A}$ for the attention scores between $\mathbf{O}_{n-A}$ and $\mathbf{O}_{n-N/A}$. Specifically, with regard to $\mathbf{O}_2$, $\mu_{2-A} < \mu_{2-N/A}$ was confirmed for all the experiments because all the p-values were $p = 1.0$ and the histogram in Figure 4.4.

Figure 4.5 shows the actual video frames in the last 10s of **GA1**. For both the heatmaps of the attention probability of $\mathbf{O}_1$ or $\mathbf{O}_2$, because there was no overlap between the view fields of the actors or object points, the attention scores of $\mathbf{O}_1$ and $\mathbf{O}_2$ were lower. However, for the heatmap of the attention probability of $\mathbf{O}_3$, there was a crowded area in the view fields of actors near the object point. Therefore, the attention scores of $\mathbf{O}_3$ were higher than those of $\mathbf{O}_1$ and $\mathbf{O}_2$.

4.3.2 Real Operating Room Surveillance Videos

4.3.2.1 Datasets

We used three videos recorded in three operating rooms, denoted as **S1**, **S2**, and **S3**, respectively. Table 4.2 shows the summary of the GA data. We visually extracted three objects that were the focus of attention in each surgery. We cropped two one-minute-long videos ($30 \text{ FPS} \times 60\text{s} = 1800 \text{ frames}$), one with the surgical team members focusing on each object and another in which they did not focus on any objects.

Figure 4.6 shows the placements for the actual operating room surveillance videos. The objects in **S1** were $\mathbf{O}_1$, a monitor under the surveillance camera; $\mathbf{O}_2$, a dissection table where STs were carried out following the excision of the affected part; and $\mathbf{O}_3$, the operating table. The objects in **S2** were $\mathbf{O}_1$, a PC placed to right-back from the surveillance camera; $\mathbf{O}_2$, the dissection table where STs were performed after excision of the affected part; $\mathbf{O}_3$, the operating table. The objects in **S3** were $\mathbf{O}_1$, a monitor

Table 4.1: P-values of one-sided Mann-Whitney U comparison test of the attention scores of $\mathbf{O}_{n-A}$ and those of $\mathbf{O}_{n-N/A}$ of **GA1**, **GA2**, and **GA3**. The null hypothesis was set at $\mu_{n-A} = \mu_{n-N/A}$, and the alternative hypothesis was set at $\mu_{n-A} > \mu_{n-N/A}$.

	$\mathbf{O}_1$	$\mathbf{O}_2$	$\mathbf{O}_3$
GA1	$p = 1.000$	$p = 1.000$	$p = 0.000$
GA2	$p = 0.769$	$p = 1.000$	$p = 0.000$
GA3	$p = 0.000$	$p = 1.000$	$p = 1.000$

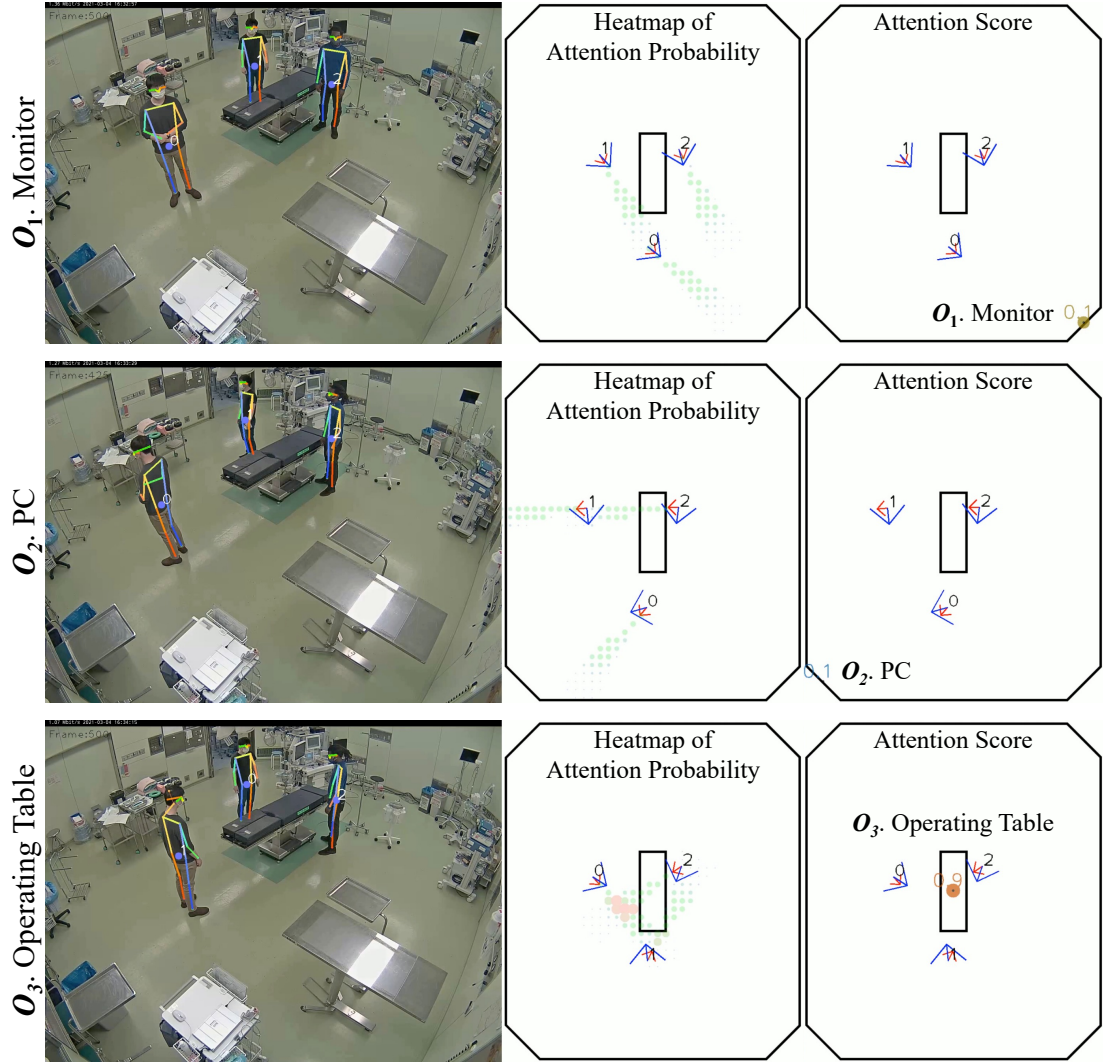


Figure 4.5: Actual video frames in last 10s of **GA1**. The top panel shows that the actors were looking at O_1 (monitor). The middle shows that the actors were looking at O_2 (PC). The bottom panel shows that the actors were looking at O_3 (Operating Table).

under the surveillance camera; O_2 , a PC placed near the monitor; and O_3 , the operating table. The videos where attention was not paid to any of the objects were cropped out of the videos of the actors preparing for surgery because they were performing various individual tasks in the operating room.

4.3.2.2 Results

Figure 4.7 shows the histograms of the attention score results. The top, middle, and bottom row shows the results for **S1**, the middle row shows the results for **S2**, and the bottom row shows the results for **S3**. The legend of “ O_n -N/A” denotes that the actors were **not** paying **attention** to O_n and “ O_n -A” indicated that they were paying **attention** to O_n .

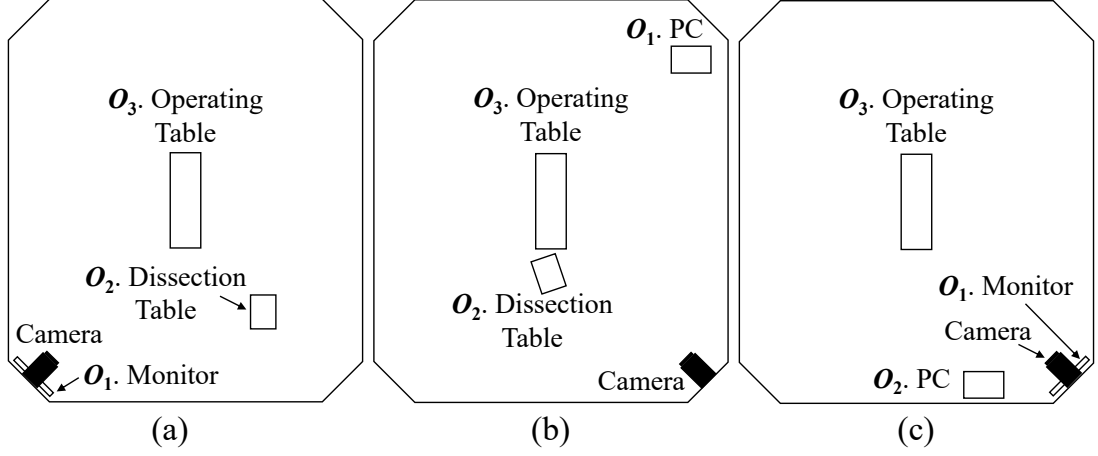


Figure 4.6: Placement of objects of focus in actual operating room surveillance videos. (a) The objects in **S1** were **O₁**, a monitor under the surveillance camera; **O₂**, a dissection table where STs were carried out following the excision of the affected part; and **O₃**, the operating table. (b) The objects in **S2** were **O₁**, a PC placed to right-back from the surveillance camera; **O₂**, the dissection table where STs were performed after excision of the affected part; **O₃**, the operating table. (c) The objects in **S3** were **O₁**, a monitor under the surveillance camera; **O₂**, a PC placed near the monitor; and **O₃**, the operating table.

Across all the cases, the results indicated that the distribution of the attention score of **O_n-A** was biased toward higher values than the attention score of **O_n-N/A** was. We set the null hypothesis as $\mu_{n-A} = \mu_{n-N/A}$, and the alternative hypothesis as $\mu_{n-A} > \mu_{n-N/A}$. We conducted the one-sided Mann-Whitney U test whose result is shown in Table 4.3. With all $p < 0.05$, there were significant differences in all the results and the attention scores of **O_n-A** were biased higher than the attention score of **O_n-N/A**.

Figure 4.8 shows the video frames for the experiment **S3** conducted to compare **O₁-A**

Table 4.2: Summary of GA data. The objects of focus in each surgery are indicated. The GA frame is the total number of frames for the group attention activities. The Non-GA frame is the total number of frames where the participants did not pay attention to any particular object. Each video was cropped to be one-minute long (30 FPS $\times$ 60s = 1800 frames).

Disposition	Object of Focus	GA Frame	Non-GA Frame
S1	Monitor	1800	1800
	Dissection Table	1800	1800
	Operating Table	1800	1800
S2	PC	1800	1800
	Dissection Table	1800	1800
	Operating Table	1800	1800
S3	Monitor	1800	1800
	PC	1800	1800
	Operating Table	1800	1800

with $\mathbf{O}_1$ -N/A. In the case in $\mathbf{O}_1$ -N/A (top), because none of the surgical team members were looking at $\mathbf{O}_1$, the heatmap of the attention probability showed nothing around $\mathbf{O}_1$. In the case in $\mathbf{O}_1$ -A (bottom), because some members gathered around $\mathbf{O}_1$ and gazed at $\mathbf{O}_1$, the heatmap of the attention probability was accumulated on $\mathbf{O}_1$. This result indicates that the proposed model was effective for recognizing the area and the objects of focus.

4.4 Discussion

When the surgical team members were not huddled together or gazing in the same direction, as in the top and middle of Figure 4.5, or they were not looking at any particular object, such as in the top of Figure 4.8, there was no crowded area of attention probability. On the other hand, when the surgical team members were converged in a spot or looking towards one particular direction, as shown at the bottom of Figure 4.5 and the bottom of Figure 4.8, there was a crowded area of attention probability. Therefore, the attention probability estimated by the proposed method was effective for predicting the gaze area of the surgical team members.

However, because we did not implement any object tracking in our experiment, the attention scores of objects were inferred from the predefined object positions. It is necessary to integrate object tracking algorithms to enable fully automated group attention recognition. This may be achieved by implementing object tracking models using neural networks [56–58] along with an optical flow for object recognition.

By investigating the experiment data, we observed that the pose estimation accuracy varied according to the individual orientation. For the situation results, refer to Section 4.3.1.2. When the surgical team members were gazing at $\mathbf{O}_1$, which was under the surveillance camera, the camera captured their faces clearly; thus, we received accurate ear and nose estimation, which led to high face orientation estimation accuracy. As can be observed in the top of Figure 4.9, on the average, the confidence scores of the estimation of the nose, left ear, and right ear points on $\mathbf{GA1} - \mathbf{O}_1$ were $\geq 90\%$.

Table 4.3: P-values of one-sided Mann-Whitney U test for the comparison of the attention scores of $\mathbf{O}_n$ -A and those of $\mathbf{O}_n$ -N/A of $\mathbf{S1}$, $\mathbf{S2}$, and $\mathbf{S3}$. The null and alternative hypotheses were set as $\mu_{n-A} = \mu_{n-N/A}$ and $\mu_{n-A} > \mu_{n-N/A}$, respectively.

	$\mathbf{O}_1$	$\mathbf{O}_2$	$\mathbf{O}_3$
$\mathbf{S1}$	$p=0.000$	$p=0.000$	$p=0.000$
$\mathbf{S2}$	$p=0.000$	$p=0.000$	$p=0.000$
$\mathbf{S3}$	$p=0.000$	$p=0.000$	$p=0.000$

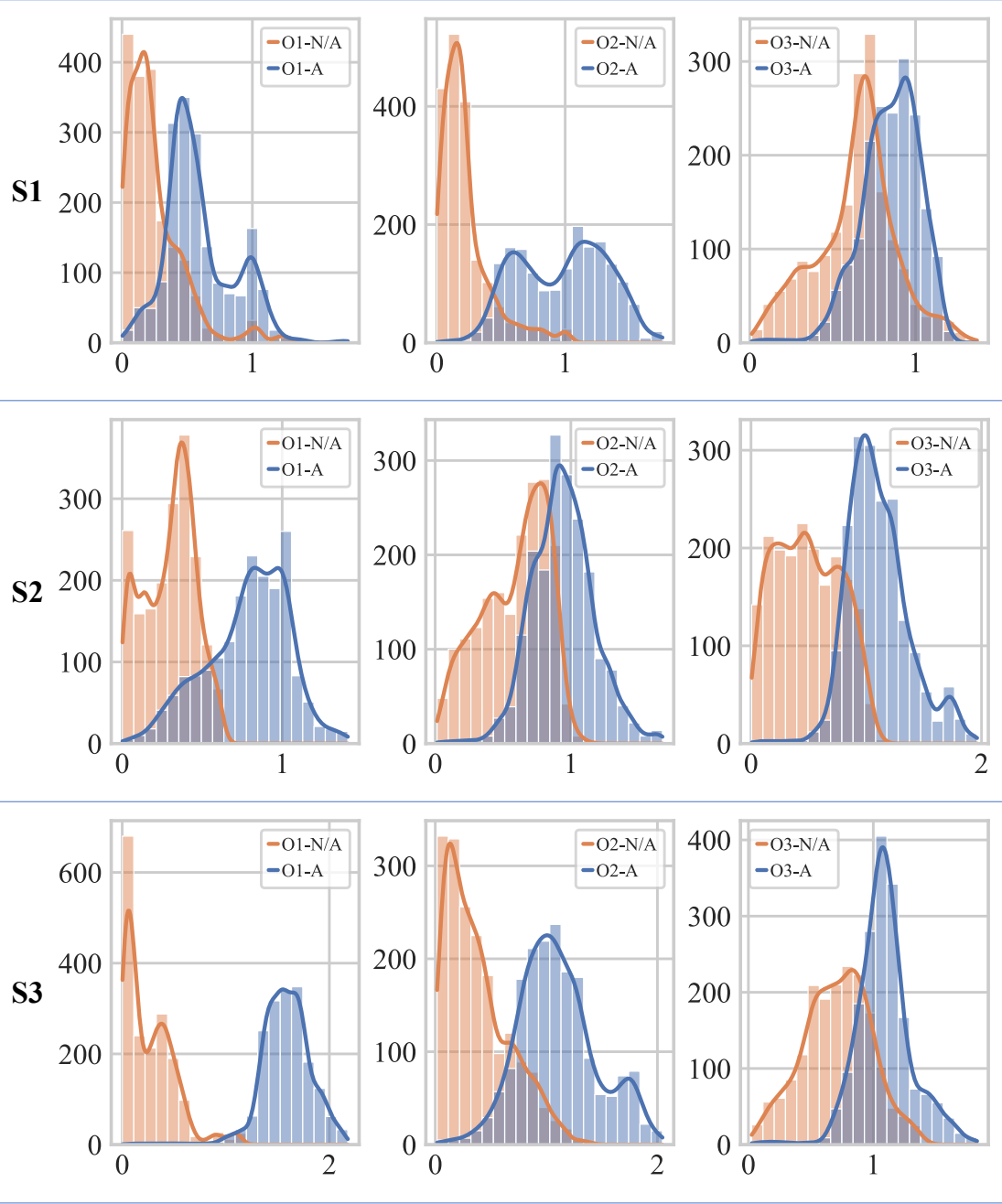


Figure 4.7: Histograms of attention score results. The top, middle, and bottom rows show the results for **S1**, **S2**, and **S3**, respectively. The legend of " O_n -N/A" denotes that the surgical team members were **not** paying **attention** to O_n and " O_n -A" denotes that they were paying **attention** to O_n .

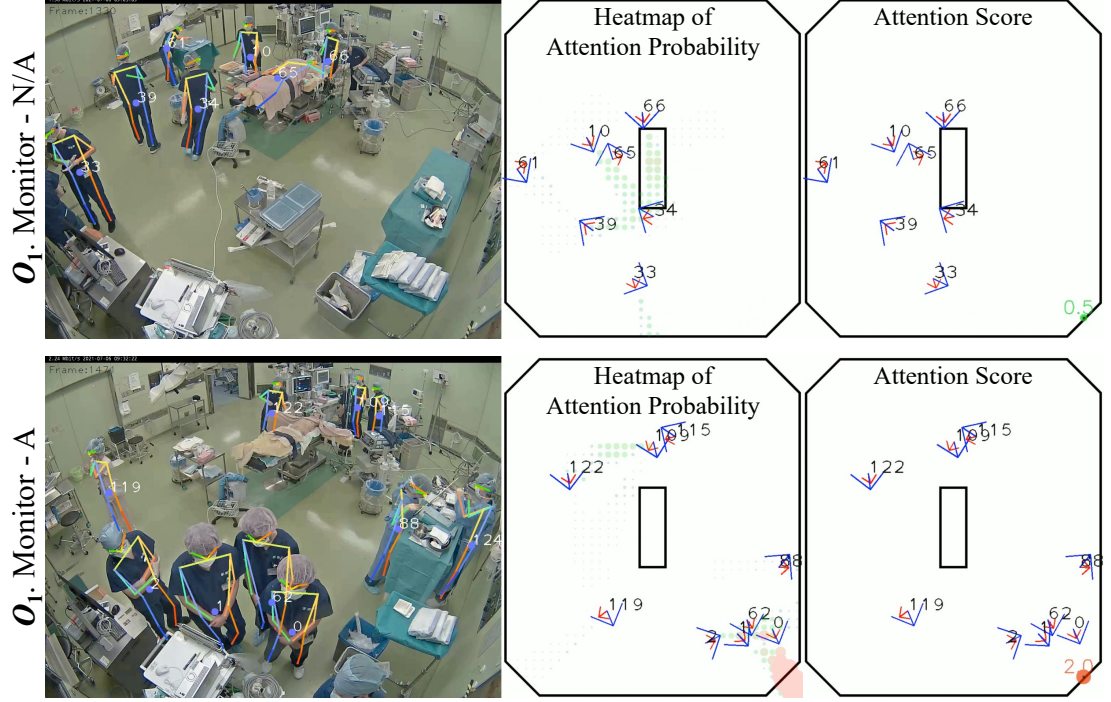


Figure 4.8: Actual video frames of **S1** - **O1**. The top shows the frame where all team members were not looking at the **O1**(monitor) and the bottom row shows the frames in which some team members gathered around the **O1**(monitor), gazing at it.

However, when the surgical team members turned their faces sideways or backwards to gaze in other directions, for example, **O2** and **O3**, the camera could not capture the facial parts clearly. The top of Figure 4.9 also shows the confidence scores of the estimation of the nose, left ear, and right ear points on **GA1** - **O2**. As can be observed, the confidence score of the right ear was lower than 80% because the right ear was occluded.

As the proposed group attention recognition model relied on individuals' face orientations, the face orientation estimation accuracy decreased when the facial features were not tracked accurately, as shown in Figure 4.10.

The performance of the proposed model can be improved by improving the face-orientation estimation. In the experiment, we used a single surveillance camera that provided only 2D information. The face orientation estimation can be improved by installing another camera to achieve 3D tracking or by applying machine-learning methods to infer 3D information from 2D images [59–63].

In addition, to estimate the viewing field of an individual, we set a threshold viewing length of 100 px and a viewing angle of 20° as the gaze region for each individual corresponding to their face orientation. However, these parameters were set by manual tuning. Thus, it is necessary to consider an optimization method.

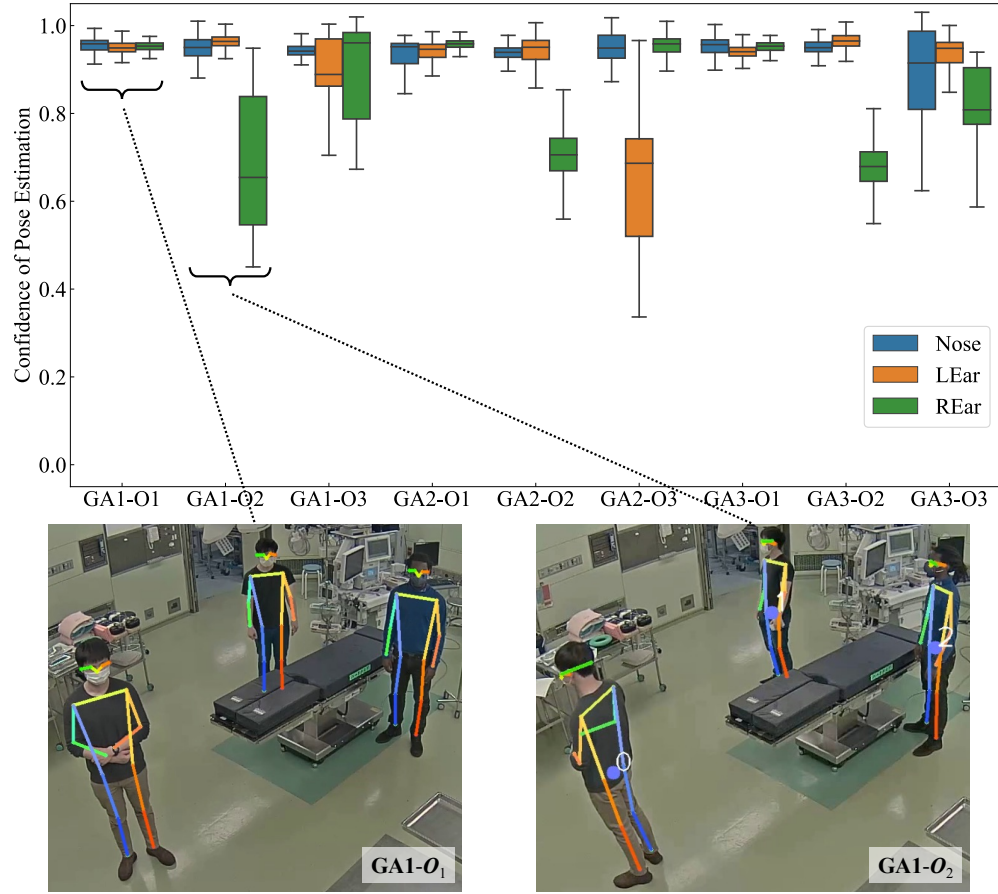


Figure 4.9: (Top) Image showing boxplots of the confidence scores of the pose estimation of the nose, left and right ear points in the simulated videos. (Bottom) The actual frames of **GA1 - O₁** and **GA1 - O₂**. For **GA1 - O₁**, because the actors were facing the surveillance camera, the confidence scores were high. However, for **GA1 - O₂**, the confidence score of the right ear was low because the actors turned sideways from the surveillance camera.

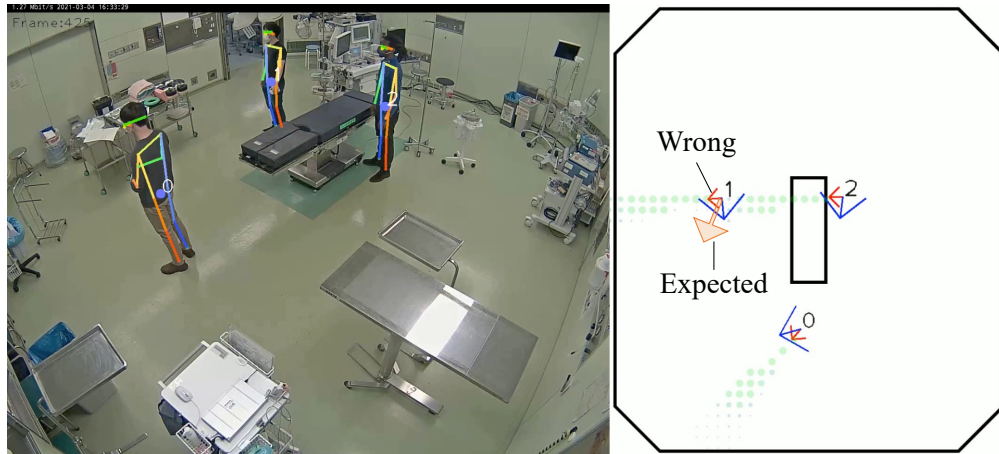


Figure 4.10: Wrong face orientation sample. The face orientation of Individual 1 (red arrow) was turned left. However, the expected face orientation is indicated by an orange arrow.

4.5 Conclusion

In this chapter, the group attention is estimated from the face orientations of surgical team members, and the areas where the surgical team members' gazes overlap are inferred. Additionally, the positions of several objects were preset, and the amount of attention each object received was quantified. From the experiments, it was found that estimating the group attention is effective for detecting the areas and objects that the surgical team members are focusing on.

Part II

End-to-End Individual Activity Recognition

In Part I, we proposed group activity recognition models that recognize interactions among surgical team members using time series keypoints of individual activities. In this part, we focus on extracting individual activity features and roles using a neural network-based activity recognition model

As mentioned in Section 1.5, intraoperative individual activities often involve standing still, with roles assigned to specific positions. This results in many ambiguous activities that are difficult to distinguish, making annotation for supervised learning challenging. To address this, we analyze individual activities captured in operating room surveillance videos by training autoencoder- and VAE-based neural network models. We use operating room surveillance video that progressed smoothly as input data for these learning models.

In Chapter 5 explores rare intraoperative activities using an anomaly detection approach. By inputting time series bounding boxes and keypoints into an autoencoder, we identify rare activities based on large reconstruction errors. Chapter 6 focuses on clustering individual activities of surgical team members using an unsupervised model. We analyze the impact of two feature types, visual and spatial, on activity clustering. Chapter 7 classifies the roles of surgical team members using semi-supervised learning. By integrating role-specific latent feature spaces into a VAE-based model, we extract activity features corresponding to each role.

Chapter 5

Unsupervised Anomaly Estimation

5.1 Introduction

In this chapter, we estimate the degree of abnormal activity. When the surgery is proceeding smoothly, the positions and poses of the surgical team members hardly change. On the other hand, when events such as emergencies occur, more people move actively, such as preparing for a transfusion.

A model trained to reconstruct individuals with little change in position or pose is used to estimate the reconstruction error as an anomaly score. The reconstruction error for surgical team members exhibiting rare activities during surgery will be large. We hypothesized that changes in the proportion of people with high reconstruction errors in the operating room could indicate events or changes in situations during the surgery.

In this chapter, we propose a model that takes the key points obtained from human bounding boxes and pose estimation as inputs and outputs reconstruction results using self-supervised learning. The model combines an autoencoder with a GAN (Generative Adversarial Network) structure. GAN is a type of generative model that can produce high-precision samples. It enhances the reconstruction results of the autoencoder, enabling highly reliable abnormal activity estimation.

5.2 Related Work

5.2.1 Reconstruction-Based Human Activity Anomaly Detection

By training the model on frequently occurring activities (normal activities), it can reconstruct them with high accuracy. When predicting rarely occurring activities (anomalous activities), the reconstruction error increases, indicating a higher degree of anomaly. Using unsupervised learning models such as autoencoders (AEs), the model learns to reconstruct only normal activities.

Gatt et al. [64] utilized Long Short-Term Memory (LSTM) and 1D convolution (1DConv)-based AE models to detect abnormal human activities, including falls. Temuroglu et al. [65] proposed a skeleton trajectory representation to handle occlusions and an AE framework for detecting abnormal pedestrian activity. Suzuki et al. [66] trained a convolutional autoencoder (CAE) based on children's good gross motor movements to detect poor limb movements as anomalies. Jiang et al. [67] presented a message-passing GRU (Gated Recurrent Unit) encoder-decoder network to detect and identify abnormal pedestrian activity in videos captured at grade crossings. The identification of anomalous pedestrians within a frame was done by identifying skeletons with reconstruction errors exceeding an empirical threshold.

AEs trained on normal data are typically expected to yield higher reconstruction errors when encountering anomalous inputs. This principle is often used as a criterion for detecting anomalies. However, this assumption does not always hold true in practice. Occasionally, AEs might successfully reconstruct the anomalies, which can lead to false negatives [68].

5.2.2 Our Motivation

During surgery proceeds smoothly, the positions and poses of the surgical team members rarely change. However, in emergencies, supporting surgical staff move significantly around the operating room. We consider that the activities within minimal changes in position and poses as normal activities. Because of their low complexity, it is possible that the model can reconstruct them with high accuracy. Conversely, activities involving significant movement around the operating room are considered anomalous, leading to higher reconstruction errors.

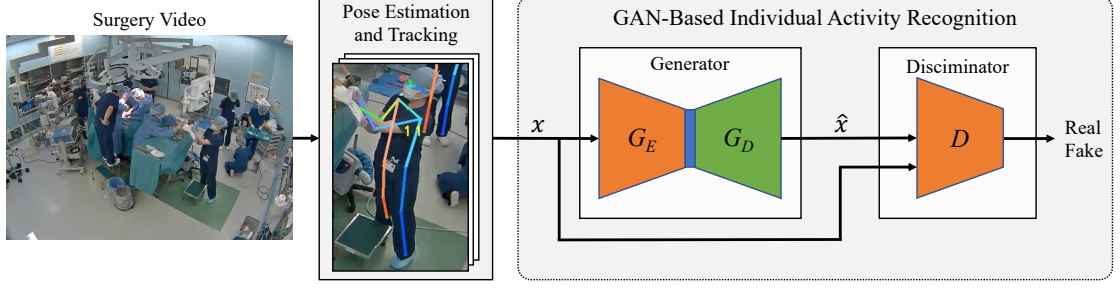


Figure 5.1: System flow comprising two components: 1) pose estimation and tracking, and 2) GAN-based individual activity recognition.

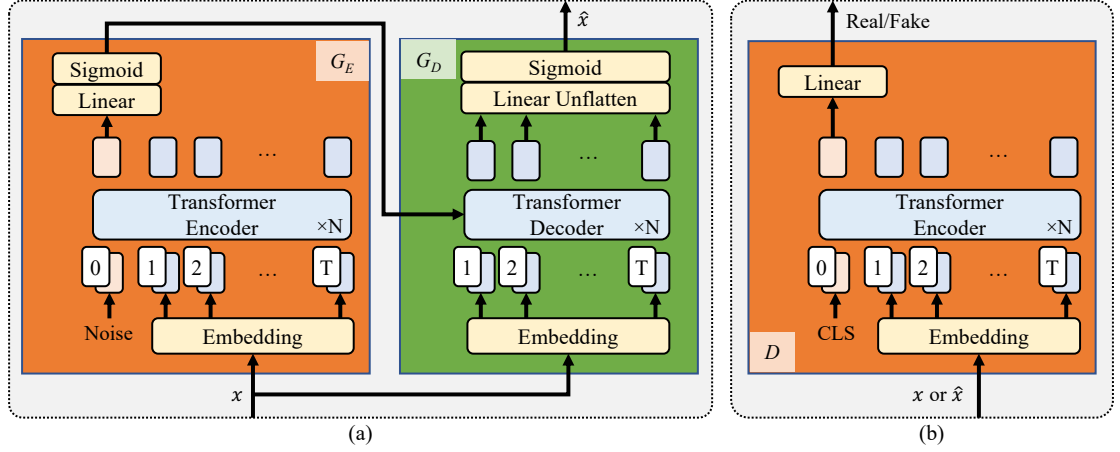


Figure 5.2: The architecture of GAN-based individual activity recognition model which consists of (a) generator and (b) discriminator.

In this chapter, we combined the structure of GAN (Generative Adversarial Networks) with AE to create a model capable of higher accuracy reconstruction. By incorporating adversarial loss from the GAN discriminator, the AE can reconstruct the data more precisely.

5.3 Methods

This section introduces the details of our method. Figure 5.1 shows system flow of proposed method which is composed of two major components: 1) pose estimation and tracking, and 2) activity anomaly estimation using time-sequential prediction.

5.3.1 Pose Estimation and Tracking

Firstly, we extract bounding boxes and key points of each individual from operating room surveillance videos as shown in Figure 5.1. We used YOLOX [69] for extracting

bounding box, High-Resolution Net (HRNet) [70] for pose estimation, and COCO [54] for the key points which includes 5 face points and 12 body points. We also used UniTrack [71] to track individuals' key points, the results are shown by yellow color numbers.

5.3.2 Activity Anomaly Estimation using Time-Sequential Prediction

We define individual activity anomaly as reconstruction error between real data and reconstructed data generated by GAN. Figure 5.2 shows the architecture of our individual activity anomaly estimation model which is similar to skip-GANomaly [72]. We customized the architecture to use the transformer layer [73] instead of a convolutional neural network.

The model is fed with time-sequential input features x into the model. The dimension of x is $(t, n, 2)$, where t is the number of frame and n is the number of points. All points are calculated in two-dimensional coordinates. The generator consisted of a transformer encoder G_E and a transformer decoder G_D , which is able to predict fake activity data $\hat{x}$. The discriminator D consisted of a transformer encoder, which catches the input x or $\hat{x}$ and predicts if the input is real or fake. G_E and D have internal parameters that are noise parameters (Noise) and class token (CLS), respectively. All of them have same embedding layers and positional embedding layers. The embedding layer consists of two pointwise convolution layers. The first one convolutes the coordinates of x to one-channel value, while the second one extracts distributed representation of x in each frame t . Positional embedding layer uses learnable absolute positional encoding [74]. Finally, we estimate individual activity anomaly scores using the mean squared error of x and $\hat{x}$.

5.4 Experiments

5.4.1 Dataset

Out of surveillance videos (30 FPS) from 6 operating rooms, we selected and cropped 40 pieces of short videos of which each long for 3 minutes. Half were used as training datasets where the surgeries proceed smoothly, and the other half were test datasets including pre- and post-operative phases, and temporary suspending.

We fed two types of 10 seconds ($t = 300$ frames) of x into our model. The first one was bounding box that was scaled by dividing by the frame size of surveillance video. The second one was key points that was substituted by the top-left corner of the

bounding box and scaled by dividing by the size of bounding box. We trained our model on each dataset.

5.4.2 Implementation

The embedding layer consists of pointwise convolution and position encoding, the embedding dimension was 256. All Transformer layers were set as same parameters, the number of attention heads was 8, the number of dimension of feed-forward network was 1024, the dropout rate was 0.1, and the activation layer was Scaled Exponential Linear Unit (SELU). The dimensions of latent features z_G and z_D were 32. We trained the models while 100 epochs with batch size of 4096. The contextual-loss was changed to Mean squared Error from skip-GANomaly [72]. The optimizer for the learning model was Adam. The learning rate was $lr = 0.001$, and betas were $(\beta_1, \beta_2) = (0.5, 0.999)$.

5.4.3 Results

The anomaly scores are shown in Figure 5.3: (a) one of training datasets which shows proceeding smoothly, and (b) one of test datasets which shows temporary suspending.

For the top of Figure 5.3 (a), there were two peaks on the orange line (tracking ID: 272), in which several activities occurred such as walking, throwing away trash, crouching, and stepping up. Figure 5.4 (a) shows the results of x and $\hat{x}$ indicated by the red arrow of Figure 5.3 (a). In this situation, the circulator did two activities: (1) changing the position of the stage while crouching, and (2) stepping up on the stage. Then, the bounding box results indicates there were large errors between real bounding boxes and fake bounding boxes. However, for the bottom of Figure 5.3 (a), all of the lines are under 0.65. The bottom of Figure 5.4 (a) indicates there were little errors between real key points and fake key points.

For the top of Figure 5.3 (b), anomaly scores of the blue line (tracking ID: 34), the orange line (tracking ID: 39), and the green line (tracking ID: 56), exceeded 0.05. In addition, for the bottom of Figure 5.3 (b), anomaly scores of the green line (tracking ID: 56), the yellow line (tracking ID: 82), and the gray line (tracking ID: 81) exceeded 0.1. Several activities also occurred at the point of high anomaly score such as walking, crouching and medical instrument passing. Figure 5.4 (b) shows the red arrow of Figure 5.3 (b) in which the circulator counted gauze while crouching. For Figure 5.3 (b), as she remained same position for 300 frames, little errors the bounding box results were occurred. However, large errors of the key points results were occurred.

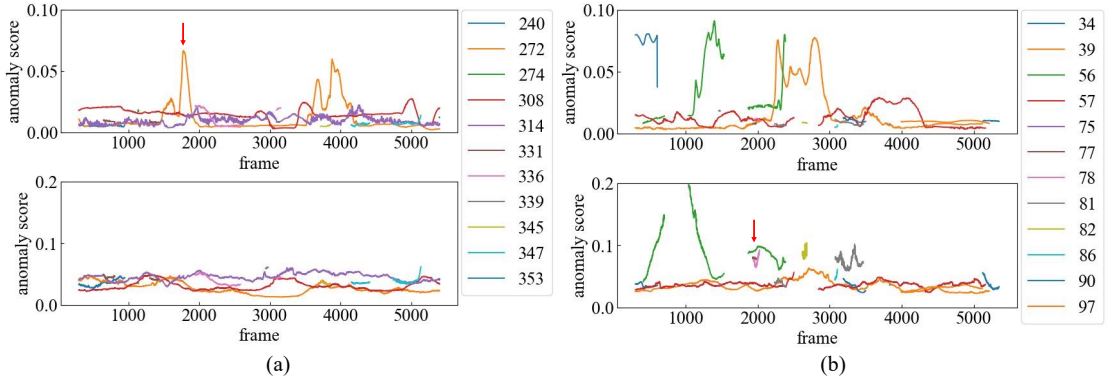


Figure 5.3: The result samples of anomaly score: (a) training dataset and (b) test dataset. The top of each result was estimated by bounding boxes. The bottom of each result was estimated by key points. The labels are tracking IDs. The red arrows show the situation of Figure 5.4 (a) and (b).

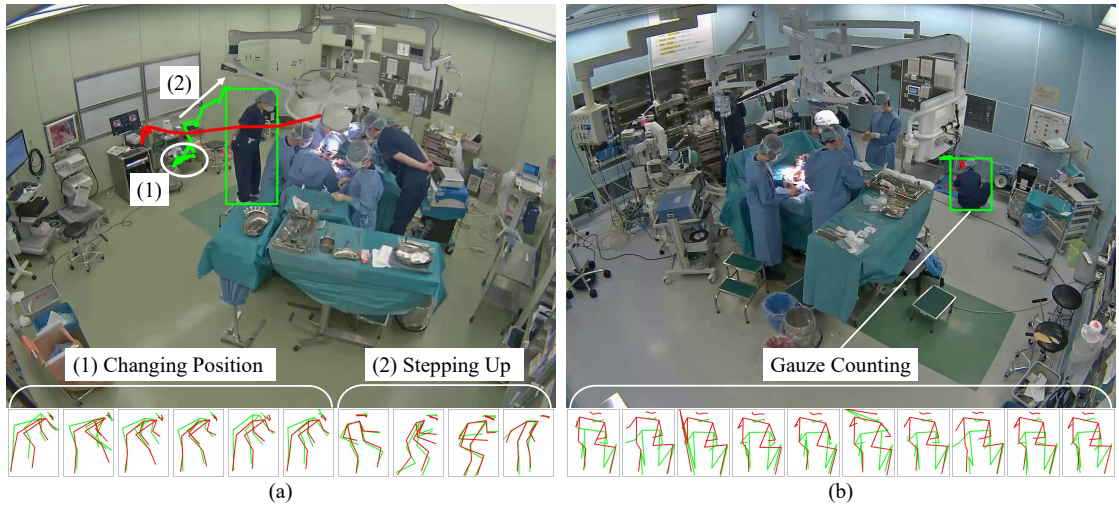


Figure 5.4: The real data and fake data indicated by the red arrows in Figure 5.3: (a) training datasets and (b) test datasets. Green and red lines show the real data and the fake data, respectively. The top of each shows the top-left corner of bounding box and the bottom of each shows key points at 30 frames (1 sec) intervals.

5.5 Discussion

When the anomaly scores from bounding boxes were observed high, large movement activities occurred such as walking around. When surgery was proceeding smoothly, the surgical team members did hardly move except for the circulator. Figure 5.3 (a) indicates that the surgery proceeded smoothly because the anomaly score of the circulator (tracking ID: 272) was high. As shown in Figure 5.3 (b), there were several high anomaly scores because surgical team members temporarily went away from the operating table. It is indicated that the prediction using bounding boxes is suitable for detecting and estimating regional movements of the individuals.

Based on the bottom of Figure 5.3 (b), we looked into the videos and found some error peaks in the anomaly scores extracted from key points. It is caused by occlusion in the videos, except for some cases where the surgical team members were clear to the camera or avoiding the occlusion occasionally, e.g. the member crouching and counting gauze in Figure 5.4. Therefore, the prediction using key points could hardly be used to detect unique activities in the surgery. We will look into masking low confidence key points to solve this issue.

In this paper, we utilized cropped videos which do not include any irregular events. In future, we will extend and improve our activity prediction to fit for detecting unexpected and urgent events by applying our model to videos including irregular issues and analyzing the anomaly scores with other latent features.

5.6 Conclusion

In this chapter, we investigated automatic analysis of intraoperative activities from surveillance videos of operating rooms. We proposed a novel method to automatically estimate individual activities using anomaly estimation based on time-sequential prediction. We evaluated the performance of the model with two types of time-sequential features: the bounding boxes and the key points. Experiment results show that the bounding boxes are suitable for predicting and detecting regional movements, while the anomaly scores using key points can hardly be used to detect activities.

Chapter 6

Unsupervised Activity Clustering

6.1 Introduction

The positions of the surgical team members are determined by their roles during the surgery. For example, the surgeon and the scrub nurse are positioned around the surgical field, and their positions do not change much during the operation. The anesthesiologist mostly stays in front of the anesthesia machine, continuously monitoring the patient. The circulator moves around a lot, carrying medical instruments in and out of the operating room and handing them to the scrub nurse. We believe that we can extract and model the activity features of each role. Furthermore, by analyzing the activity changes of surgical team members based on their roles, we can detect incidents during surgery, recognize situations such as busyness, and contribute to improving the safety and efficiency of surgeries.

However, surgeries require surgical team members to stand for prolonged periods or perform minimal but different activities based on the role of each member, which makes discerning their activity challenging. Annotating such data requires input from experienced surgical personnel, resulting in high costs. Furthermore, because surgical procedures typically last for multiple hours, manual visual annotation is not feasible [42].

To tackle this issue, we propose a novel unsupervised model that enables activity classification for long videos such as operating room videos, without the need of annotation. Our model inputs video frames and optical flow into an autoencoder using I3D [32], and extracts visual features of individual activities with RoIAlign [75]. Our model also takes advantages of spatial features which we consider crucial for recognizing human activities. In surgical rooms, there are distinct spatial characteristics, with individuals positioned around the surgical table. For example, surgeons and scrub nurses are at the

operating table, anesthetists are positioned at the head of the patient, and circulators move freely in and out of the operating room. In our model, we enhance the recognition accuracy by incorporating spatial features specific to surgical environments. As a simple feature, we manually specify the central coordinates of the operating tables and calculated the L2 norm between each center of individual’s bounding box and the center of the operating table.

6.2 Related Work

6.2.1 Deep Clustering

Numerous studies have focused on unsupervised learning methods for image classification. With the advancement of neural networks, methods, such as deep autoencoder-based (DAE-based) [76–81], variational autoencoder-based (VAE-based) [82–85], generative adversarial network-based (GAN-based) [86–88], and graph neural network-based (GNN-based) [89–91], can achieve superior accuracy in image classification compared to that of classical approaches. In DAE-based methods, deep neural networks composed of autoencoders are used to extract features and perform clustering. These methods have two variants, namely methods incorporating clustering algorithms such as k-means alongside the autoencoder [76–78] and methods simultaneously learn dedicated clustering layers [79–81]. VAE-based methods optimize the probability distribution of latent features. In most models, optimal solutions related to the evidence lower bound (ELBO) are used [92]. GAN-based methods involve adversarial training of a generator and discriminator. These techniques can also be used in clustering. The categorical GAN (CatGAN) [86] is notable in distinguishing fake data and data belonging to specific categories. GNN-based models associate feature vectors with graph vertices and learn similarity between vertices. Data with similar features exhibit proximity, resulting in close distances between vertices. In GNNs, structural information is captured from graphs, enabling intrinsic learning of data relationships.

These methods offer diverse approaches to unsupervised image classification, each with unique strengths and applications.

6.2.2 Our Motivation

Previous approaches to HAR have relied on supervised learning or semi-supervised learning using datasets such as the collective activity [93] dataset and volleyball dataset [22]. However in the context of real-world operating room videos, the traditional supervised

learning is not well-suited. Surgeries vary based on techniques, surgical team members, and operating room equipment, making robust learning difficult. Additionally, the movements of surgical team members often involve prolonged periods of standing, which makes them challenging to discern. Furthermore, accurately annotating long-term surgeries via manual observation is a formidable task. Consequently, annotating these videos requires the expertise of experienced surgeons, resulting in high annotation costs.

The objective of our study is to achieve activity recognition through completely unsupervised learning without annotation. While various deep clustering methods exist for unsupervised clustering of video sets, the target operating room video analysis requires clustering of the individuals appearing in each frame. Therefore, directly applying existing HAR models and deep clustering models to the scenarios discussed is still challenging.

Furthermore, existing HAR methods have primarily used visual features extracted by I3D. However, spatial context has been ignored in the previous studies [33–36]. Depending on the video domain (e.g., sports), the arrangement of individuals within the frame can be predetermined by their roles in the team, rendering spatial features crucial. Surgical team members are also arranged by their roles. For example, surgeons and scrub nurse are placed around the operating table, anesthetist is placed in front of the anesthesia machine, and circulator moves around the operating room. To address this gap in literature, our method combines visual features obtained using I3D [32] and RoIAlign [75] with semi-manually extracted spatial features. We then use a centroid-based model to obtain individual features and perform clustering.

6.3 Methods

6.3.1 Model Architecture

Figure 6.1 shows the architecture of the proposed model. The model consists of a) two autoencoders $\mathbf{AE}_{\phi, frame}$ and $\mathbf{AE}_{\phi, flow}$, and b) clustering layer $\mathbf{CL}_{\theta}$.

6.3.1.1 Autoencoder

The top autoencoder $\mathbf{AE}_{\phi, frame}$ for the RGB frame inputs the RGB frames $\mathbf{X}_{frame} \in \mathbf{R}^{T \times w \times h \times d}$, and outputs the reconstructed RGB frames $\hat{\mathbf{X}}_{frame} \in \mathbf{R}^{T \times w \times h \times d}$, where T denotes the number of frames, w denotes the width of frames, h denotes the height of frames, d denotes the dimension of frames, and ϕ denotes the learnable parameters.

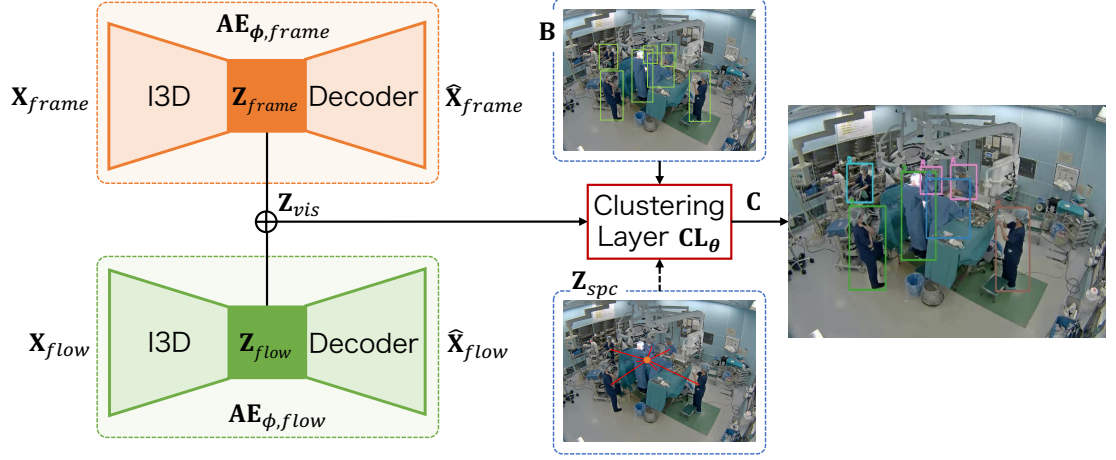


Figure 6.1: Overview of the proposed model. Visual features are extracted from two autoencoders (RGB frame, optical flow). In the clustering layer, individual features are extracted from the given bounding boxes, and spatial features from the center of the operating table and the center are added. Finally, the cluster number of each individual is output.

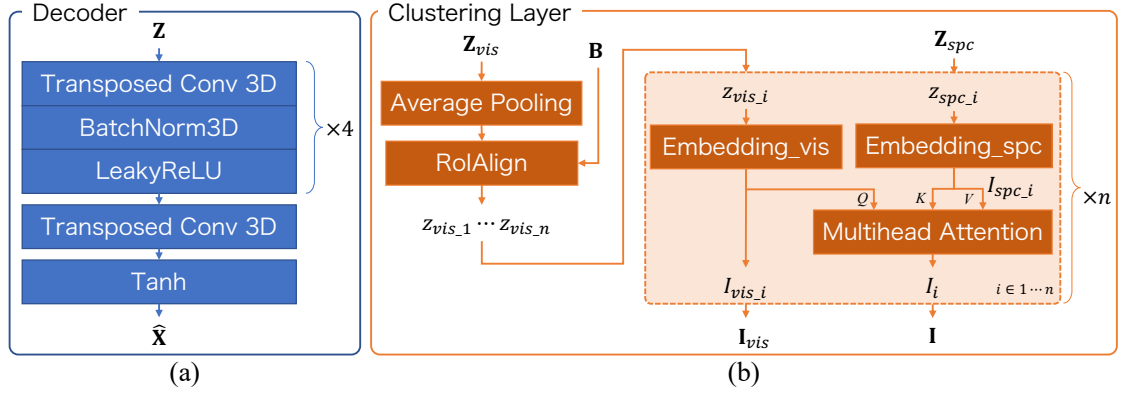


Figure 6.2: Model architectures of (a) Decoders and (b) Clustering Layers.

Similarly, the bottom autoencoder $AE_{\phi,flow}$ for the optical flow inputs the optical flows $X_{flow} \in \mathbf{R}^{T \times w \times h \times d}$, and outputs the reconstructed optical flows $\hat{X}_{flow} \in \mathbf{R}^{T \times w \times h \times d}$. We used Gunnar Farneback's algorithm [94] for computing optical flows. The parameters are set as follows: scale=0.5, level=3, window size=15, iterations=3, number of poly=5, sigma of poly=1.2.

Each autoencoder differs only in the input/output dimension (RGB frame: $d = 3$, optical flow: $d = 2$). The encoder consists of up to Mixed_3c layers of I3D. The decoder consists of four layers combining 3D transposed convolution, batch normalization, and leaky ReLU, 3D transposed convolution and tanh layers, as shown in Figure 6.2 (a).

The two autoencoders extract RGB frame feature $Z_{frame} \in \mathbf{R}^{T_z \times w_z \times h_z \times d_z}$ and optical flow feature $Z_{flow} \in \mathbf{R}^{T_z \times w_z \times h_z \times d_z}$ respectively, where T_z denotes the length of I3D

outputs, w_z denotes the width of I3D outputs, h_z denotes the height of I3D outputs, and d_z denotes the dimension of I3D outputs (for "Mixed_3c", d_z is fixed at 480). The visual feature $\mathbf{Z}_{vis}$ is obtained by adding features $\mathbf{Z}_{frame}$ and $\mathbf{Z}_{flow}$.

6.3.1.2 Clustering Layer

Figure 6.2 (b) shows the details of the clustering layer $\mathbf{CL}_\theta$, where θ denotes the learnable parameters. First, the average pooling layer is applied to $\mathbf{Z}_{vis}$ in the time axis direction. Second, RoIAlign is applied to $\mathbf{Z}_{vis}$ to extract the features $z_{vis,1} \cdots z_{vis,n}$ in the bounding boxes $\mathbf{B} \in \mathbf{R}^{2 \times 2 \times n}$, where n denotes the number of individuals in the frame. $\mathbf{B}$ consists of top-left corner and bottom-right corner points of individuals and is extracted on the last frame of $\mathbf{X}_{frame}$ by YOLOv8 [95]. Third, for each $z_{vis,i} (i \in 1 \cdots n)$, embedding_vis layer is applied to embed individual visual feature $I_{vis,i} \in \mathbf{R}^{d_i}$, and $\mathbf{I}_{vis}$ is obtained by collecting $I_{vis,1} \cdots I_{vis,n}$, where d_i denotes the dimension of individual feature and $i \in 1 \cdots n$.

The L2 norm between the center of the operating table and the center of each individual bounding box are input as $\mathbf{Z}_{spc}$ to add spatial feature. First, Embedding_spc layer is applied to embed individual spatial feature $z_{spc,i}$ to get $I_{spc,i} \in \mathbf{R}^{d_i}$. Second, $I_{vis,i}$ is inputted as query and $I_{spc,i}$ is inputted as key and value into the multihead attention layer to fusion visual feature and spatial feature to obtain I_i . $\mathbf{I}$ is obtained by collecting I_i .

We train our model by two patterns; a) using only visual feature $I_{vis,i}$ and b) using both visual and spatial features I_i .

6.3.2 Clustering Algorithm

We considered the model with reference to Xie et al. [79]. The proposed model has k cluster centroids $\mu \in \mathbf{R}^{k \times d_i}$. We calculated the similarity probability of I_i and μ_j as soft distribution $s_{ij} \in \mathbf{S}$ by using student's T-distribution as follows:

$$s_{ij} = \frac{(1 + \|I_i - \mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_j (1 + \|I_i - \mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}, \quad (6.1)$$

where i denotes the index of individuals, j denotes the index of cluster centroids, and α denotes the freedom of student's T-distribution. Target distribution $t_{ij} \in \mathbf{T}$ is defined as follows:

$$t_{ij} = \frac{s_{ij}^2 / \sum_i s_{ij}}{\sum_j (s_{ij}^2 / \sum_i s_{ij})}, \quad (6.2)$$

where t_{ij} is defined by s_{ij} . By learning the model to match $\mathbf{S}$ with $\mathbf{T}$, the proposed model is trained to cluster individuals. Finally, clustering results $c_i \in \mathbf{C}$ is defined as follows:

$$c_i = \arg \max_j s_{ij} \quad (6.3)$$

6.3.3 Loss Functions

Reconstructive loss functions are defined as the mean squared error (MSE) of the input and output images of the respective autoencoder. The RGB frame loss function lr_{frame} is calculated by the MSE between input RGB frames $\mathbf{X}_{frame}$ and reconstructed RGB frames $\hat{\mathbf{X}}_{frame}$. The optical flow loss function lr_{flow} is calculated by the MSE between input optical flows $\mathbf{X}_{flow}$ and reconstructed optical flows $\hat{\mathbf{X}}_{flow}$.

$$lr_{frame} = \text{MSE}(\mathbf{X}_{frame}, \hat{\mathbf{X}}_{frame}) \quad (6.4)$$

$$lr_{flow} = \text{MSE}(\mathbf{X}_{flow}, \hat{\mathbf{X}}_{flow}) \quad (6.5)$$

Clustering loss lc is calculated by the KL-divergence between soft distribution $\mathbf{S}$ and target distribution $\mathbf{T}$.

$$lc = \text{KL}(\mathbf{S}||\mathbf{T}) \quad (6.6)$$

6.3.4 Training Algorithm

First, autoencoders are pretrained to stabilize learning by Algorithm 1. Second, encoder, decoder, and clustering layer are trained by using various loss functions by Algorithm 2.

6.4 Experiments and Results

6.4.1 Dataset

The dataset was created from surveillance camera installed in operating rooms of Kyoto University Hospital. Table 6.1 presents the contents of the dataset. The training data

Algorithm 1 Pretraining Autoencoders

```

1: Inputs:  $(\mathbf{X}_{frame}, \mathbf{X}_{flow})$ 
2: Outputs: None
3: Training loop:
4: while epoch < 200 do
5:   1. Forward autoencoders:
6:    $\hat{\mathbf{X}}_{frame} \leftarrow \mathbf{AE}_{\phi, frame}(\mathbf{X}_{frame})$ 
7:    $\hat{\mathbf{X}}_{flow} \leftarrow \mathbf{AE}_{\phi, flow}(\mathbf{X}_{flow})$ 
8:
9:   2. Calculate losses:
10:   $lr_{frame} = \text{MSE}(\mathbf{X}_{frame}, \hat{\mathbf{X}}_{frame})$ 
11:   $lr_{flow} = \text{MSE}(\mathbf{X}_{flow}, \hat{\mathbf{X}}_{flow})$ 
12:
13:  3. Backward losses and update model parameters  $\phi$ :
14:   $lr_{frame} \rightarrow \mathbf{AE}_{\phi, frame}$ 
15:   $lr_{flow} \rightarrow \mathbf{AE}_{\phi, flow}$ 
16: end while

```

Algorithm 2 Training Entire Model

```

1: Inputs:  $(\mathbf{X}_{frame}, \mathbf{X}_{flow}, \mathbf{Z}_{spc}, \mathbf{B})$ 
2: Outputs: None
3: Training loop:
4: while epoch < 100 do
5:   1. Forward autoencoders:
6:    $(\hat{\mathbf{X}}_{frame}, \mathbf{Z}_{frame}) \leftarrow \mathbf{AE}_{\phi, frame}(\mathbf{X}_{frame})$ 
7:    $(\hat{\mathbf{X}}_{flow}, \mathbf{Z}_{flow}) \leftarrow \mathbf{AE}_{\phi, flow}(\mathbf{X}_{flow})$ 
8:    $\mathbf{Z}_{vis} = \mathbf{Z}_{frame} + \mathbf{Z}_{flow}$ 
9:
10:  2. Clustering:
11:  if use spacial feature  $\mathbf{Z}_{spc}$  then
12:     $\mathbf{I} \leftarrow \mathbf{CL}_{\theta}(\mathbf{Z}_{vis}, \mathbf{Z}_{spc}, \mathbf{B})$ 
13:    Update soft distribution  $\mathbf{S} \leftarrow (\mathbf{I}, \mu)$  by Equation 6.1
14:  else
15:     $\mathbf{I}_{vis} \leftarrow \mathbf{CL}_{\theta}(\mathbf{Z}_{vis}, \mathbf{B})$ 
16:    Update soft distribution  $\mathbf{S} \leftarrow (\mathbf{I}_{vis}, \mu)$  by Equation 6.1
17:  end if
18:  Update target distribution  $\mathbf{T} \leftarrow \mathbf{S}$  by Equation 6.2
19:  Get clustering result  $\mathbf{C} \leftarrow \mathbf{S}$  by Equation 6.3
20:
21:  3. Calculate losses:
22:   $lr_{frame} = \text{MSE}(\mathbf{X}_{frame}, \hat{\mathbf{X}}_{frame})$ 
23:   $lr_{flow} = \text{MSE}(\mathbf{X}_{flow}, \hat{\mathbf{X}}_{flow})$ 
24:   $lc = \text{KL}(\mathbf{S} || \mathbf{T})$ 
25:
26:  4. Backward losses and update model parameters  $\phi, \theta$ :
27:   $lr_{frame} + lc \rightarrow \text{encoder of } \mathbf{AE}_{\phi, frame}$ 
28:   $lr_{flow} + lc \rightarrow \text{encoder of } \mathbf{AE}_{\phi, flow}$ 
29:   $lr_{frame} \rightarrow \text{decoder of } \mathbf{AE}_{\phi, frame}$ 
30:   $lr_{flow} \rightarrow \text{decoder of } \mathbf{AE}_{\phi, flow}$ 
31:   $lc \rightarrow (\mathbf{CL}_{\theta}, \mu)$ 
32: end while

```

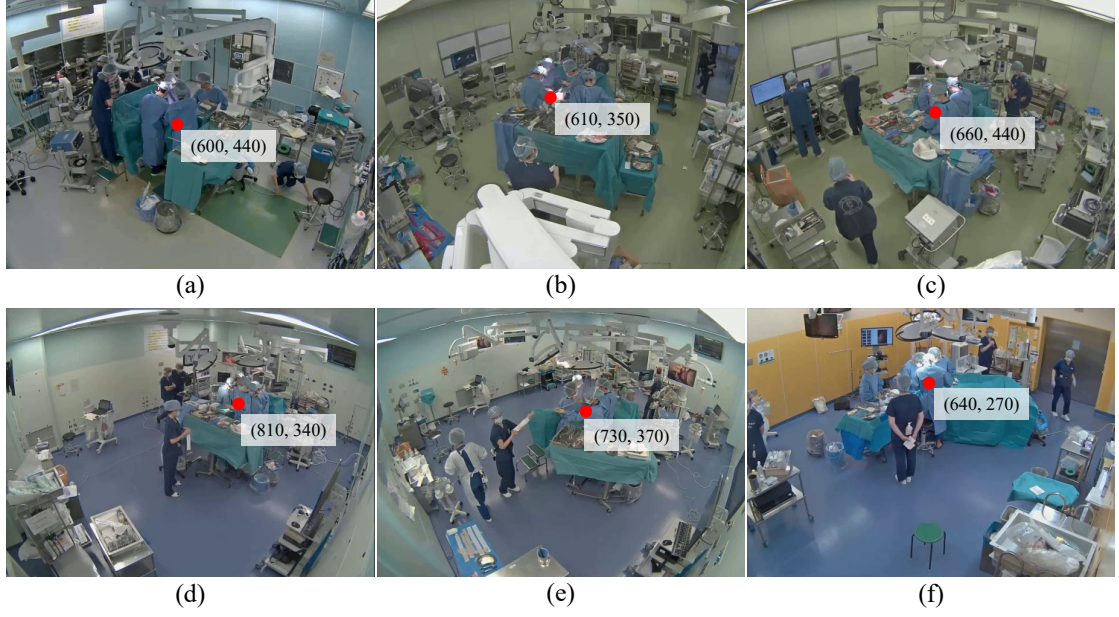


Figure 6.3: Sample frames of operating room videos. (a) to (f) corresponding in order from $S1$ to $S6$. Each red point is manually pointed to the center of operating table in each surgery.

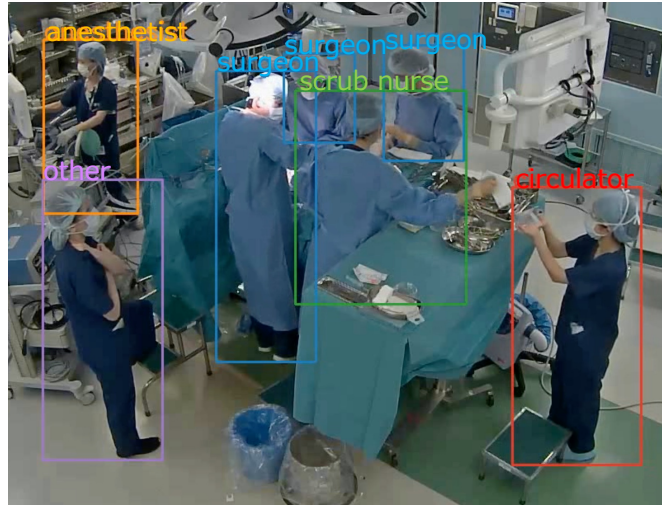


Figure 6.4: Label samples of the dataset that includes surgeon, anesthetist, scrub nurse, circulator, and other.

consisted of 20 operating room videos from three surgeries $S1 \dots S3$. The test data consisted of 9 operating room videos from six operating rooms $S1 \dots S6$, and 9 of the videos were the same surgeries of the training data. Figure 6.3 shows the sample RGB frames of all surgeries. The FPS of all videos were 30 and each was 5 second long (150 frames).

We set the five labels surgeon, anesthetist, scrub nurse, circulator and other. Figure 6.4 shows the samples of all labels.

6.4.2 Implementation

6.4.2.1 Model Architecture

We used up to "Mixed_3c" layers of I3D as encoders in both autoencoders. We constructed the decoders in the both autoencoders, as shown in Figure 6.2 (a). Table 6.2 shows the structure of the 3D transposed convolution layers, and the negative slope of the all Leaky ReLU layers were fixed as 0.1.

Figure 6.2 (b) shows the structure of the clustering layer. We set the output size of RoIAlign to 5 and set output dimensions of both embedding_vis layer and embedding_spc layer to 32. The number of heads and the dropout rate of the multihead attention layer were set to 4 and 0.1, respectively. We set the dimensions of the cluster centroids to 32 to match the outputs of embedding layers. α of student's T-distribution was fixed to 2.

Table 6.1: Dataset contents. Training dataset consists of 20 videos from three surgeries $S1 \cdots S3$. Test dataset consists of nine videos from six surgeries $S1 \cdots S6$ (three of videos are the same surgeries of training dataset).

Dataset	$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	Total
Training	7	6	7	N/A	N/A	N/A	20
Test	2	2	2	1	1	1	9

Table 6.2: Structure of 3D transposed convolution layers in the decoder.

Layer No	Filter Size	Stride	Padding
1	(4, 4, 4)	(2, 2, 2)	(1, 1, 1)
2	(4, 4, 1)	(2, 2, 1)	(1, 1, 0)
3	(4, 4, 1)	(2, 2, 1)	(1, 1, 0)
4	(1, 1, 1)	(1, 1, 1)	(0, 0, 0)
5	(1, 1, 1)	(1, 1, 1)	(0, 0, 0)

Table 6.3: Conditions of training.

No	Cluster Number k	with Spatial Feature
1	5	False
2	5	True
3	10	False
4	10	True

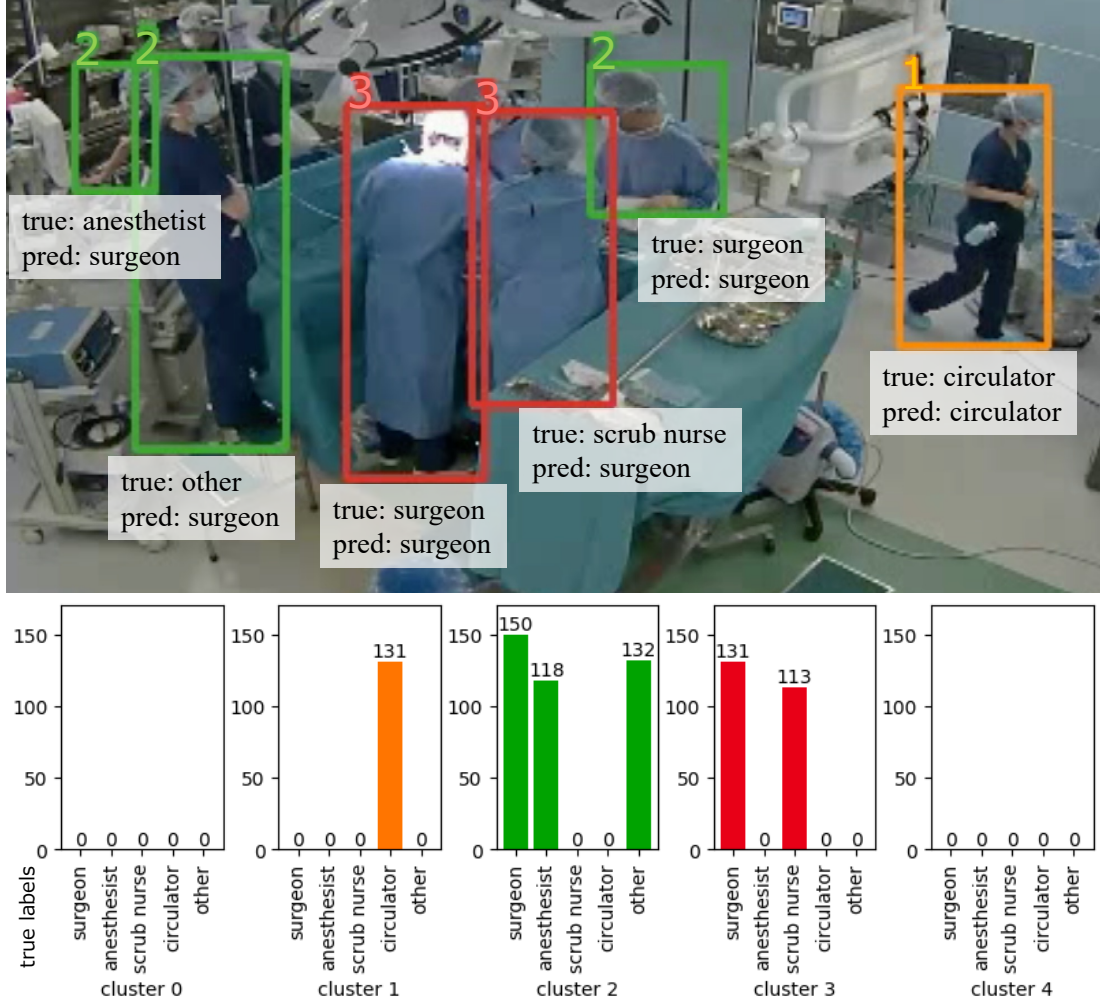


Figure 6.5: Example of linking clustering results to true labels. Cluster 1 (orange) was linked to circulator, cluster 2 (green) was linked to surgeon, and cluster 3 (red) was linked to surgeon.

6.4.2.2 Dataset and Features

Human Detection We detected human bounding boxes $\mathbf{B}$ using YOLOv8. To achieve unsupervised learning, we used YOLOv8 which was trained on the COCO dataset[96] without fine-tuning.

Visual Feature Extraction We used the frames $\mathbf{X}_{frame}$ and optical flows $\mathbf{X}_{flow}$ by resizing from 1280×940 to 640×360 . We resized bounding boxes $\mathbf{B}$ by the ratio as match as resized frames.

Spatial Feature Extraction We manually set the center of operating rooms as shown in Figure 6.3. Next, we calculated L2 norms between the centers of operating rooms and the centers of individual bounding boxes as $\mathbf{Z}_{spc}$. Finally, we resized spatial features $\mathbf{Z}_{spc}$ by the ratio as match as resized frames.

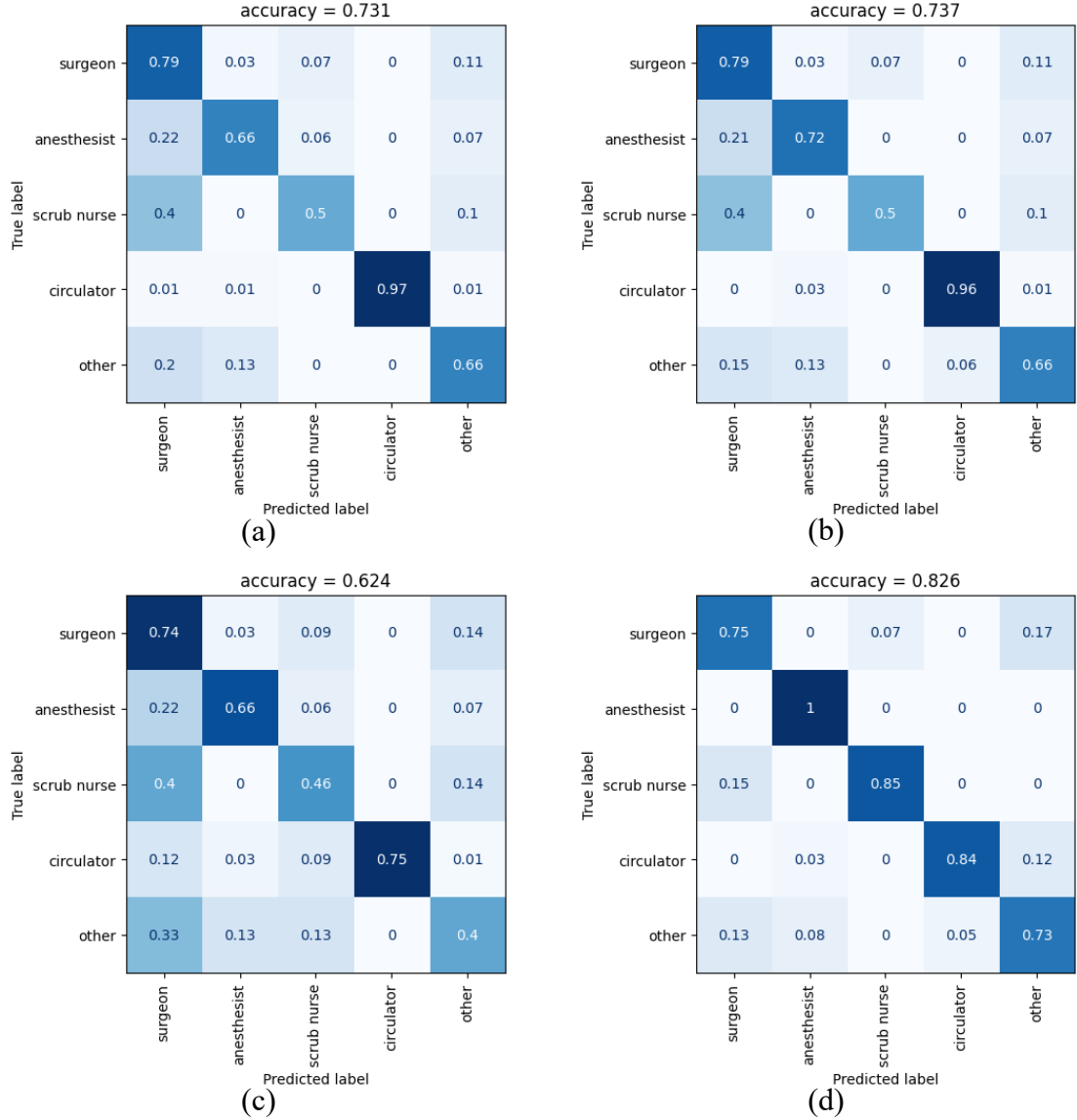


Figure 6.6: Confusion matrix of the training dataset. (a) to (d) correspond in order from condition 1 to 4 of Table 6.3. The labels are set as follows; 0: surgeon, 1: anesthetist, 2: scrub nurse, 3: circulator, 4: other.

6.4.2.3 Training

For pretraining autoencoders, we set the number of epochs to 200. We used Adam for optimizing, and the learning rates of both were 0.0001 and (β_1, β_2) were (0.9, 0.999). Both learning rates were decreased by 0.1 times by epoch = 100.

For training all structures of the model, we set the number of epochs to 100. We used Adam for optimizing both encoders, both decoders and clustering layer and cluster centroids, and the learning rates of all were 0.0001 and (β_1, β_2) were (0.9, 0.999). All of the learning rates were decreased by 0.1 times by epoch = 50.

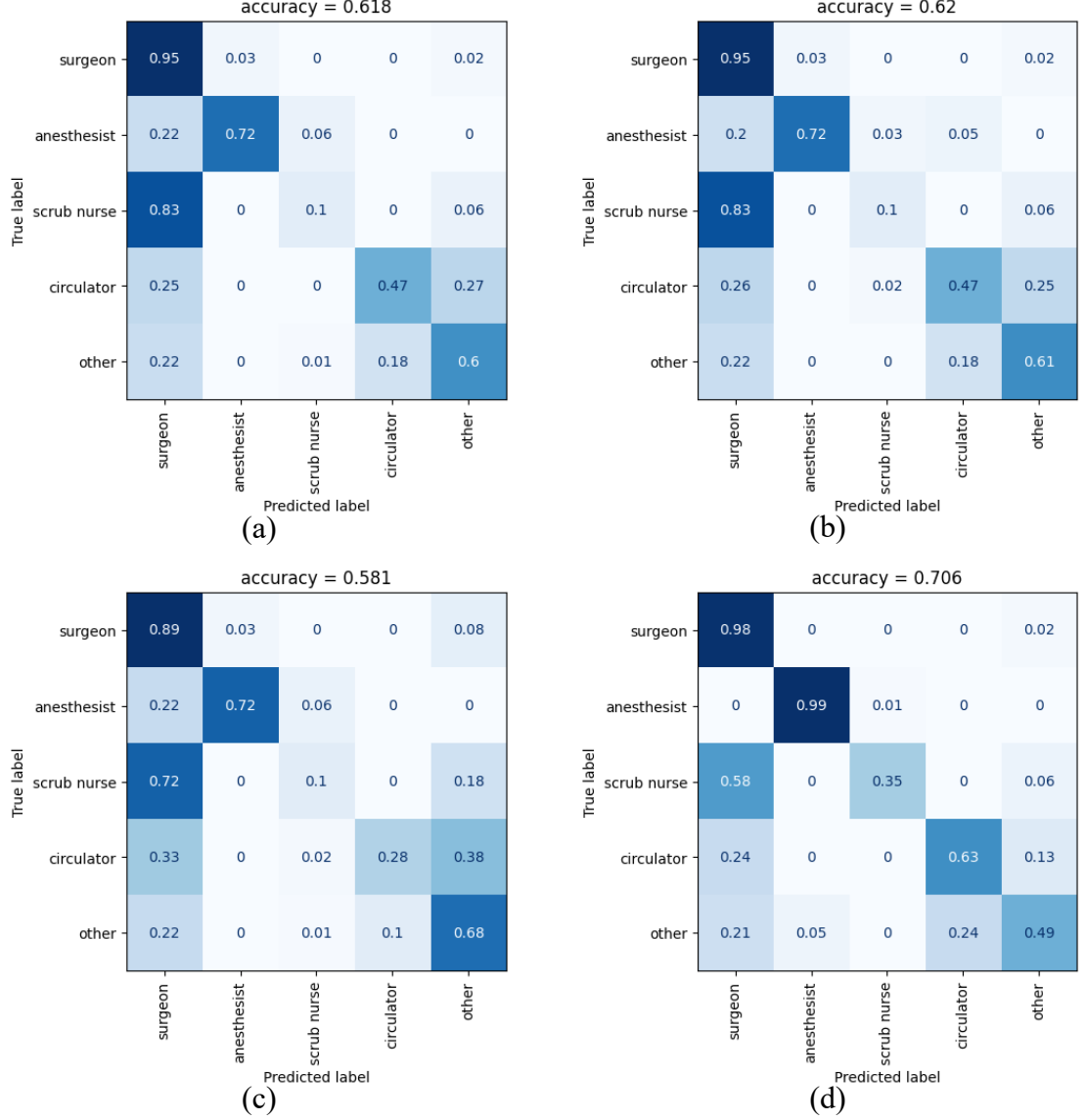


Figure 6.7: Confusion matrix of the test dataset. (a) to (d) correspond in order from condition 1 to 4 of Table 6.3. The labels are set as follows; 0: surgeon, 1: anesthetist, 2: scrub nurse, 3: circulator, 4: other.

We trained our model under the conditions shown in Table 6.3. We varied the number of clusters k , and whether to use spatial features $\mathbf{Z}_{spc}$ to analyze the effects of parameter changes.

Activities of surgical members are vary significantly depending on the roles. For example, surgeons are positioned around the operating table and rarely change their position, while circulators move around the operating room to transport medical instruments. To match the number of roles in the surgical team defined in Section 6.4.1, we set $k = 5$. However, due to variations in the arrangement of medical instruments and

background colors in the operating room, the view from the operating room surveillance video differs. Therefore, we increased the number of clusters to $k = 10$

6.4.3 Results

For an evaluation, linking the clustering results to the true labels of the dataset is necessary. For each video, predicted labels are selected by the most frequent true labels of each clustering results. Figure 6.5 shows an example of label linking. For instance, the highest number of true labels in cluster 2 (green) was “surgeon” at 150, and the predicted labels of all individuals in cluster 2 was set as “surgeon”.

Training Dataset

Figure 6.6 shows the accuracy and confusion matrix of the training dataset. The results correspond to (a) in order from model condition No.1 in Table 6.3. Focusing on Figure 6.6 (a) and (b), both accuracies were approximately 0.73 and the confusion matrix were almost the same when $k = 5$. Focusing on Figure 6.6 (c), the accuracy decreased to approximately 0.62 when spatial feature was not used with $k = 10$. By contrast, focusing on Figure 6.6 (d), the accuracy increased to approximately 0.83 when the spatial feature was used with $k = 10$. In particular, the accuracy of anesthetist and scrub nurse improved significantly over the other results by 0.85. However, only the accuracy of the circulator decreased by approximately 0.1 compared with the case in which spatial features were not used.

Test Dataset

Figure 6.7 shows the accuracy and confusion matrix of the test dataset. Similar to the training dataset, the results correspond to (a) in order from model condition No.1 in Table 6.3. As shown in Figure 6.7 (a),(b) and (d), the accuracy decreased the accuracy approximately 0.1 and for (c), the accuracy decreased approximately 0.05 compared with the results of the training dataset. For all results, errors that the true label of anesthesiologist was predicted to surgeon, with (a) and (b) were approximately 0.83 and (c) was approximately 0.72 and (d) was approximately 0.58. Numerous errors were observed in the circulator and other labels for all the confusion matrices. Circulator accuracy was approximately 0.47, and the accuracy of others was approximately 0.6 for (a) and (b); the circulator accuracy was approximately 0.28 and the accuracy of other was approximately 0.68 for (c); circulator accuracy was approximately 0.63 and the accuracy of other was approximately 0.49 for (d).

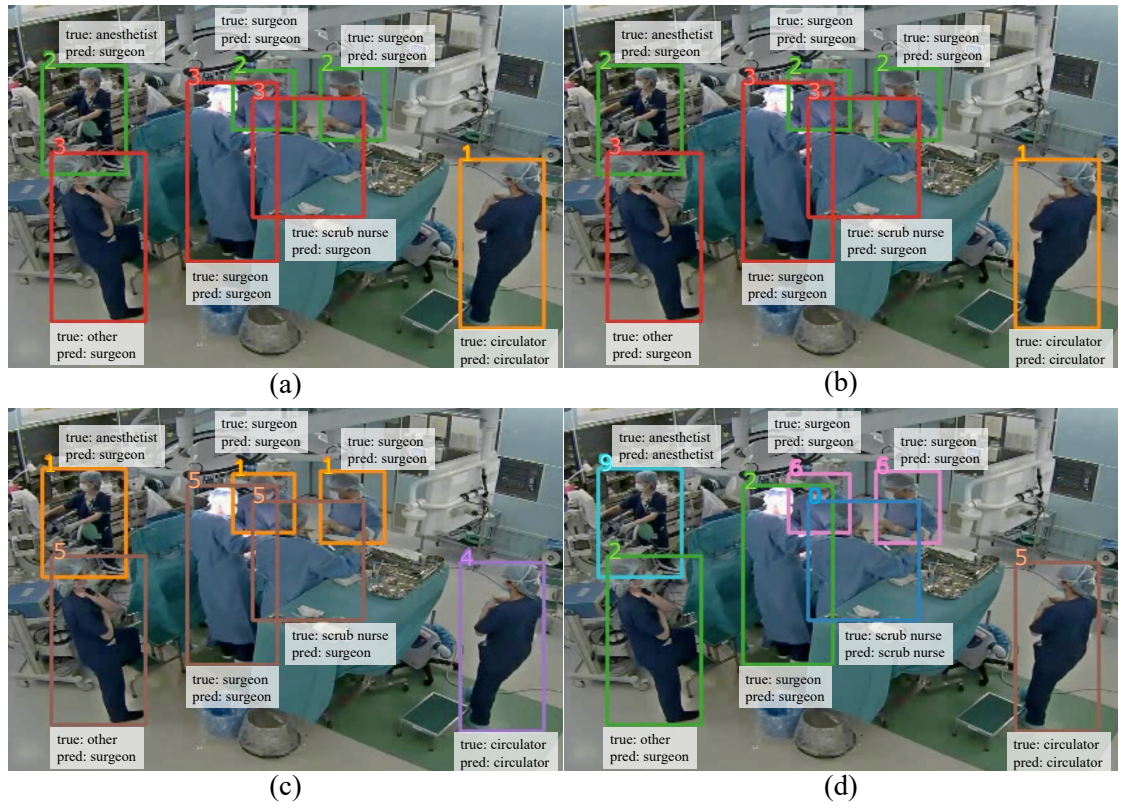


Figure 6.8: Clustering result samples of the training dataset. (a) to (d) correspond in order from condition 1 to 4 of Table 6.3. The true label and predicted label are shown in each white text box.

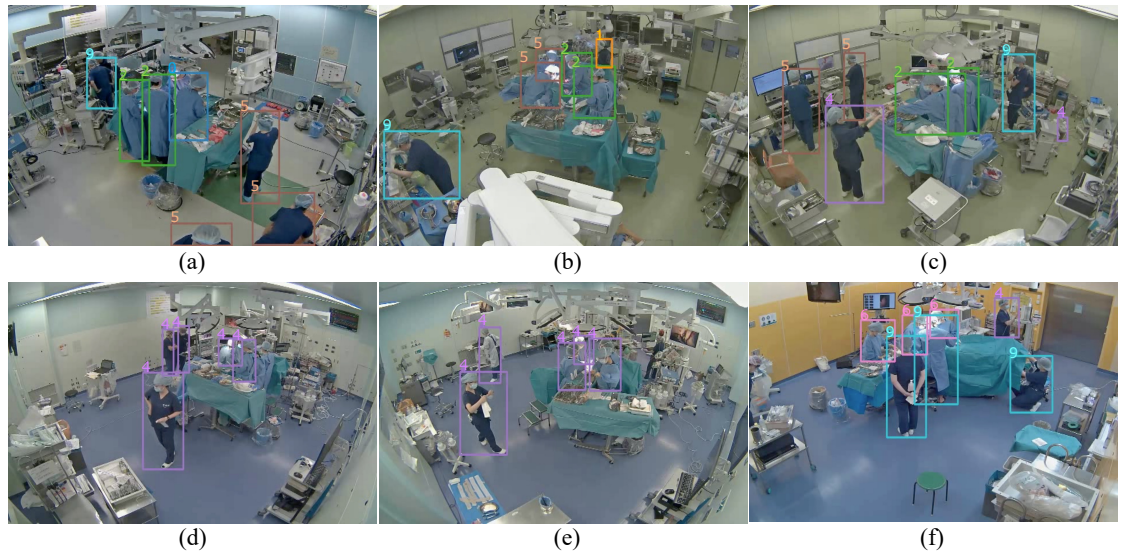


Figure 6.9: Clustering result of the test dataset using the model of condition 4 of Table 6.3. (a) to (f) correspond in order from $S1$ to $S6$.

6.5 Discussion

6.5.1 Implication

Figure 6.6 reveals that the results for the training dataset across all models accurately detected surgeons and circulators with a detection rate exceeding 0.7. Figure 6.8 depicts result samples from the training dataset. Circulators positioned on the right of each image often moving around the operating room to delivery medical equipment and tend to be higher optical flows compared with others. Therefore, we consider that circulators were more easily classified in one cluster.

By contrast, an operating room has multiple surgeons, and when clusters and labels are linked using the algorithm described at the beginning of Section 6.4.3, clusters tend to be determined to be surgeons. Therefore, when the predicted label was a surgeon, the false detection rate for other labels increased. In Figure 6.8 (a),(b), clusters 2 and 3 were identified as surgeons, while in (c), clusters 1 and 5 are identified as surgeons, and in (d), clusters 2 and 6 are identified as surgeons. The scrub nurse working next to the surgeon is particularly noteworthy; the probability of the predicted label being a surgeon exceeds 0.4 in (a), (b), and (c), and resulting in an scrub nurse detection rate of approximately 0.5.

For Figure 6.8 (d), cluster 0 was identified as an scrub nurse, classified into a various cluster than surgeons, which resulted in a low false detection rate and an scrub nurse detection rate of 0.85.

For Figure 6.7, across all models, the results for the test dataset consistently achieved an accurate detection of surgeons and anesthetists, exceeding a detection rate of 0.7. However, similar to the training dataset, when the predicted label was a surgeon, a high probability of results were classified as other labels. Figure 6.9 shows results samples from the model under condition 4 of Table 6.3. Figure 6.9 (a)···(c) are the results of surgeries $S1 \cdots S3$, which confirm that the clusters could be classified as well as the training dataset.

Figure 6.9 (d)···(f) are the results of surgeries $S4 \cdots S6$, which are not included in the training dataset, and the clustering was hardly working. All individuals in Figure 6.9 (d) and (e) were classified as 4 so that everyone was predicted as surgeon using the algorithm mentioned in Section 6.4.3. Because the proposed model used autoencoder to perform feature extraction, feature extraction was not successful using operating room videos that were not included in the training data.

We consider that by learning different models for each operating room, accurate clustering can be achieved in any operating room. The proposed model can be applied simply by inputting the operating room surveillance video and the coordination of the operating table in each operating room, thus requiring little effort to learn different models for each operating room.

6.5.2 Effect of Spatial Feature

In the proposed model, the L2 norm between manually specified center of the operating table and the centers of individual bounding boxes were added as spatial features.

In condition 4 of Table 6.3, the number of clusters was set at double the number of labels $k = 10$, and the accuracy was improved by 0.1 by adding spatial features. Figure 6.8 (d) reveals that scrub nurse in cluster 0 and anesthetist in cluster 9 were classified into different clusters from surgeons. Figure 6.3 shows that scrub nurses are positioned nearly at the same location as the center of the operating table, whereas anesthetists were positioned away from the centers of the operating tables. Therefore, scrub nurses and anesthetists were classified into separate clusters from surgeons.

6.5.3 Limitation

In the proposed method, YOLOv8 was used for extracting individual bounding boxes. Consequently, recognizing all individuals was not possible, and several objects were recognized as individuals. For Figure 6.9, the coordinates of individual bounding boxes could be misaligned, or objects were misclassified as individuals indicated in the rightmost bounding box in Figure 6.9 (c). In this study, YOLOv8 was not fine-tuned for operating room videos to achieve unsupervised learning. Instead, a model pretrained on the COCO dataset [96] was used. We considered that tuning YOLOv8 for each type of video could lead to accurate recognition of individual bounding boxes, improving clustering accuracy.

6.6 Conclusion

In this chapter, an unsupervised learning model was applied to operating room videos for activity classification. The results revealed that the proposed model could perform activity classification with an accuracy of more than 0.7 on the training dataset and more than 0.6 on the test dataset. The accuracy was improved by changing the number of clusters k and adding spatial features unique to operating rooms. Under condition 4,

we could classify the activities with an accuracy of approximately 0.83 on the training dataset and approximately 0.71 on the test dataset.

Chapter 7

Semi-Supervised Role Classification

7.1 Introduction

In this chapter, we propose a model that leverages the advantages of the models introduced in Chapter 5 and Chapter 6 to classify roles using semi-supervised learning. Similar to model in Chapter 5, our proposed model is trained using bounding boxes and pose estimation keypoints as inputs. In Chapter 6, pixel values were used; however, they lack spatial features and include background information, which may lead to the extraction of non-activity-related features. In contrast, bounding boxes and pose estimation keypoints consist of coordinate points that preserve spatial motion features while eliminating background noise. Additionally, compared to pixel values, they require less data, improving both model training efficiency and prediction accuracy.

Unlike the models in Chapter 5 and Chapter 6, the proposed model performs activity clustering using a single model, enabling simultaneous activity reconstruction and classification. During training, we use partially labeled role data for semi-supervised learning. In Chapter 6, the unsupervised clustering model often misclassified similar roles, such as the surgeon and the scrub nurse, due to their overlapping activities in surveillance video. By incorporating a small amount of labeled data, our proposed model significantly improves classification accuracy. Additionally, since annotating full-length surgical videos is resource-intensive, our approach enables high-accuracy role classification using only a partially annotated dataset.

Our model is based on variational autoencoders (VAE) [97], which maximize likelihood to enable sampling from the latent variable space, representing key data features.

However, since many surgical activities involve standing still, the input skeleton models often appear similar, increasing the risk of mode collapse. To address this, we incorporate a vector quantized variational autoencoder (VQ-VAE) [98], which discretizes latent variables using vector quantization, preventing mode collapse and improving reconstruction accuracy, and a stochastically quantized variational autoencoder (SQ-VAE) [99], which enhances latent space utilization by stochastically discretizing latent variables.

These VAE models represent all activity classes within a shared latent space, which may fail to capture differences between roles. To overcome this limitation, we propose a categorical SQ-VAE (CSQ-VAE) where each role has a distinct latent variable distribution, facilitating more precise representation of role-specific activity features.

7.2 Theoretical Background

7.2.1 VAE

Consider the observed data distribution $x \in \mathbb{R}^D$ and its empirical distribution $p_d(x) \in \mathbb{R}^D$. A VAE consists of the three main components; an encoder, which maps the input data x to the latent variable $z \in \mathbb{R}^{d_z}$ via the approximate the posterior $q_\varphi(z|x)$; a latent variable space, which represents the prior distribution of z denoted as $p(z)$; and a decoder, which approximate the prior $p_\theta(x|z)$ reconstructing the observed data x from z . Here, D and d_z represent the dimensions of the observed data and latent variable, respectively, while φ and θ denote the trainable parameters.

The objective of VAE is to maximize the Evidence Lower Bound (ELBO) on the marginal log-likelihood.

$$\mathcal{L} = \mathbb{E}_{q_\varphi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\varphi(z|x)||p(z)) \quad (7.1)$$

The prior distribution $q_\varphi(x|z)$ and the posterior distribution $p_\theta(x|z)$ are represented by normal distributions with diagonal covariance.

$$q_\varphi(x|z) \sim \mathcal{N}(g_\varphi(x), \sigma_\varphi(x)\mathbf{I}) \quad (7.2)$$

$$p_\theta(x|z) \sim \mathcal{N}(f_\theta(z), \sigma^2\mathbf{I}), \quad (7.3)$$

where $g_\varphi : \mathbb{R}^D \rightarrow \mathbb{R}^{d_z}$, $\sigma_\varphi : \mathbb{R}^D \rightarrow \mathbb{R}^{d_z}$ and $f_\theta : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^D$ are mapping functions.

The first term in Equation 7.1 represents the negative reconstruction error, which is minimized using the mean squared error. The second term is the regularization term,

which approximates the posterior distribution to the prior distribution of the latent variable. To enable gradient-based optimization, the reparametrization trick is used, expressing z as a differentiable function of $\epsilon \sim \mathcal{N}(0, 1)$.

$$z = \mu_\varphi(x) + \sigma_\varphi(x) \odot \epsilon \quad (7.4)$$

7.2.2 Categorical VAE

Categorical VAE [100, 101] replaces the latent variable space $p(z)$ with $p(c)p(z|c)$ to represent different discrete latent spaces for each category, where $c \in 1, \dots, C$ and C denotes the number of categories. And, the encoder $q_\varphi(z|x)$ is replaced with $q_\varphi(z|x, c)$ by adding the category label as a condition discriminated in $q_\varphi(c|x)$. The decoder $p_\theta(x|z)$ is also replaced with $p_\theta(x|z, c)$ to add the category label as a condition.

The objective of categorical VAE is to optimize the learnable parameters φ and θ by maximizing the ELBO on the marginal log-likelihood. Due to the introduction of the categorical condition c [101], Equation 7.1 is modified because of adding condition c as follows:

$$\begin{aligned} \mathcal{L} = & \mathbb{E}_{q_\varphi(z, c|x)} [\log p_\theta(x|z, c)] \\ & - \mathbb{E}_{q_\varphi(c|x)} [D_{KL}(q_\varphi(z|x, c) || p_\theta(z|c))] - D_{KL}(q_\varphi(c|x) || p(c)) \end{aligned} \quad (7.5)$$

Semi-supervised learning can utilize both labeled and unlabeled data [100]. A challenge in handling unlabeled data is that the categorical distribution cannot be reparametrized. Kingma et al. [102] address this issue by marginalizing over all possible categories c as follows:

$$\mathcal{U} = \sum_c q_\varphi(c|x) \mathcal{L} + \mathcal{H}(q_\varphi(c|x)), \quad (7.6)$$

where $\mathcal{H}$ denotes the entropy.

Finally, the ELBO for semi-supervised learning is obtained by the following equation.

$$\mathcal{J} = \mathcal{L} + \mathcal{U} + \alpha \cdot \log q_\varphi(c|x), \quad (7.7)$$

where α is the scalar balance between the generative and discriminative objectives.

This method faces the problem of increased computational complexity as the number of categories C grows. To address this, the Gumbel-Softmax [100] introduces a

continuous distribution for categorical variables, enabling efficient sampling. Additionally, it allows parameter gradients to be computed using the reparameterization trick, facilitating end-to-end learning.

7.2.3 VQ-VAE

The VQ-VAE [98] constructs a learnable latent embedding space $e \in \mathbb{R}^{K,d}$, where K is the number of embedding vectors, and d_z is their dimensionality. The model quantizes the latent variables deterministically using vector quantization. For a given input x the encoder produces an output $z_e \in \mathbb{R}^{n,d_z} = q_\varphi(x)$, where n represents the number of latent variables. The discrete latent variable $z_i \in \mathbb{R}^{d_z}, i \in 1, \dots, n$ is then quantized by performing a nearest neighbor in the embedding space e . The posterior distribution $q(z_i|x)$ is defined as a one-hot probability expressed as follows:

$$q(z_i|x) = \begin{cases} 1 & \text{for } k = \operatorname{argmin}_j \|z_e - e_j\|_2, \\ 0 & \text{otherwise} \end{cases} \quad (7.8)$$

z_e is mapped to the closest latent embedding variable e in the quantization process, forming the decoder input $z_q(x)$, as follows:

$$z_q(x) = e_k, \text{ where } k = \operatorname{argmin}_j \|z_e - e_j\|_2 \quad (7.9)$$

The ELBO of VQ-VAE is as follows:

$$\mathcal{L}_{VQ} = \log p(x|z_q(x)) + \|\operatorname{sg}[z_e] - e\|_2^2 + \beta \|z_e - \operatorname{sg}[e]\|_2^2, \quad (7.10)$$

where $\operatorname{sg}[\cdot]$ denotes the stop-gradient operator. Although β depends on the scale of the reconstruction loss, it does not significantly affect within the results range of 0.1 and 2.0 [98].

7.2.4 SQ-VAE

SQ-VAE [99] employs a stochastic vector quantization process for latent variables using a self-annealing method applied to a trainable categorical posterior distribution. This approach eliminates the need for empirical hyperparameter tuning.

Similar to VQ-VAE, SQ-VAE features a latent embedding space e . Given an input x , the encoder produces an output $z_e \sim q_\varphi(z_e|x)$ which is then transformed into a continuous latent variable z through the dequantization process $p_\varphi(z|z_e)$. The probabilistic

quantization process $\hat{p}_\varphi(z_q|z) \sim p_\varphi(z|z_q)p(z_q)$ is derived from Bayes' theorem, where $p(z_q)$ represents the prior distribution of the discrete latent space e . By stacking the dequantization process $p_\varphi(z|z_e)$ with the probabilistic quantization process $\hat{p}_\varphi(z_q|z)$, z_e and z_q are learned to approximate each other.

The ELBO of SQ-VAE is given by:

$$\begin{aligned} \mathcal{L}_{SQ} = & \mathbb{E}_{q_\omega(z|x)\hat{p}_\varphi(z_q|z)} \left[\log \frac{p_\theta(x|z_q)p_\varphi(z|z_q)}{q_\omega(z|x)} \right] \\ & + \mathbb{E}_{q_\omega(z|x)} \mathcal{H}(\hat{p}_\varphi(z_q|z)) + \text{const}, \end{aligned} \quad (7.11)$$

where $q_\omega(z|x) = p_\varphi(z|z_e)q_\phi(z_e|x)$ with $\omega = \{\phi, \varphi\}$. The categorical distribution $\hat{p}_\varphi(z_q|z)$ is approximated using the Gumbel-Softmax method, and the gradients are optimized end-to-end training.

If x is continuous, the dequantization process and the decoder assume a Gaussian distribution.

$$p_\varphi(z|z_q) = \mathcal{N}(z_q, \Sigma_\varphi), \quad (7.12)$$

where $\Sigma_\varphi = \sigma_\varphi \mathbf{I}$, and

$$p_\theta(x|z_q) = \mathcal{N}(f_\theta(z_q), \sigma \mathbf{I}), \quad (7.13)$$

where $\sigma \in \mathbb{R}^+$ and $f_\theta : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^D$. Furthermore, by Bayes' theorem, the probabilistic quantization process is formulated as the Mahalanobis distance between of e and $z_{q,i}$ with the variance Σ_φ .

$$\hat{p}_\varphi(z_{q,i} = e_k|z) = \text{softmax}_k \left(\left\{ -\frac{(e_j - z_i)^T \Sigma_\varphi^{-1} (e_j - z_i)}{2} \right\}_{j=1}^K \right) \quad (7.14)$$

Here, Σ_φ is a trainable parameter that controls the degree of quantization probability during training. As σ^2 gradually decreases, Σ_φ becomes less stochastic, leading to the quantization process to converge toward deterministic quantization (self-annealing) [99].

Finally, the objective function of SQ-VAE is given by:

$$\begin{aligned} \mathcal{L}_{SQ} = & \mathbb{E}_{q_\omega(z|x)\hat{p}_\varphi(z_q|z)} \left[\frac{1}{2\sigma^2} \|x - f_\theta(z)\|_2^2 + \frac{1}{2\sigma_\varphi^2} \|z - e\|_2^2 \right] \\ & + \mathbb{E}_{q_\omega(z|x)} \mathcal{H}(\hat{p}(z_q|z)) + \text{const} \end{aligned} \quad (7.15)$$

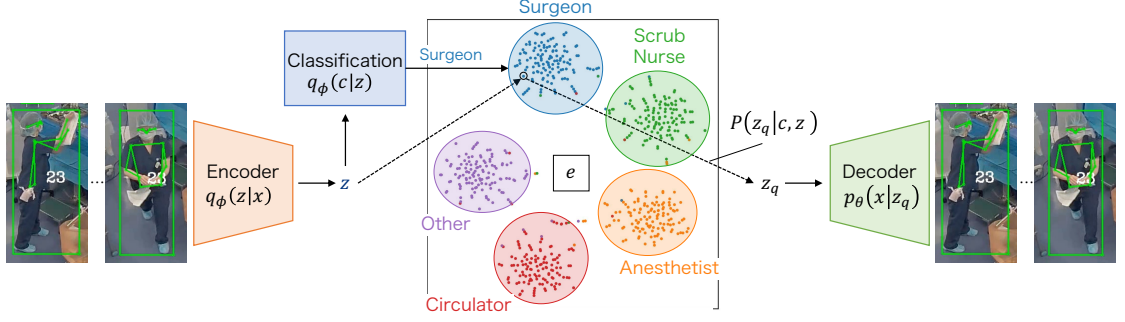


Figure 7.1: Overview of Categorical SQ-VAE, which consists of Encoder, Decoder, and Classification Head, and the latent variable spaces for each label.

7.3 Categorical SQ-VAE

SQ-VAE has high representation capacity in the latent variable space and strong reconstruction ability. However, it does not support data classification. To address this limitation, we introduce CSQ-VAE, an extension of SQ-VAE that incorporates a classification process. An overview of CSQ-VAE is shown in Figure 7.1.

CSQ-VAE features discrete latent spaces $e \in \mathbb{R}^{C,K}$ that vary by category, where C represents the number of categories and K denotes the size of each discrete latent space to capture distinct features for its respective category. Additionally, by conditioning on the category, CSQ-VAE can generate samples that exhibit the prominent characteristics of each category from the discrete latent space.

Consider an observed data distribution $x \in \mathbb{R}^D$ with its empirical distribution $p_d(x) \in \mathbb{R}^D$. The encoder consists of an inference network $q_\phi(z|x)$, which maps x to a continuous latent variable z . The classification layer includes a discriminative network $q_\phi(c|z)$. The probabilistic quantization process $p_\phi(z_q|z, c)$ selects z_q as the element from the discrete latent space e_c that is closest to z in terms of the Mahalanobis distance. The decoder consists of a generative network $p_\theta(x|z_q)$.

The ELBO of the categorical SQ-VAE is given by:

$$\begin{aligned}
\log p_\theta(x, c) &\geq \mathcal{L}_{csqvae} = \\
&\mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z, c)} \left[\log \frac{p_\theta(x|z_q)p_\varphi(z|z_q)p(z_q|c)p(c)}{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z, c)} \right] \\
&= \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z, c)} \left[\log \frac{p_\theta(x|z_q)p_\varphi(z|z_q)}{q_\phi(z|x)} \right] \\
&+ \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)} \mathcal{H}(\hat{p}_\varphi(z_q|z, c)) \\
&+ \mathbb{E}_{q_\phi(z|x)} \left[D_{KL}(q_\varphi(c|z)||p(c)) \right] + \text{const}
\end{aligned} \tag{7.16}$$

The categorical distribution $\hat{p}_\varphi(z_q|z, c)$ and $q_\varphi(c|x)$ are approximated using the Gumbel-Softmax method, and the gradients are optimized end-to-end training.

In this study, x is continuous as it consists of the bounding box and keypoints of the skeleton model.

$$p_\varphi(z_i|z_q) = \mathcal{N}(z_{q,i}, \Sigma_\varphi), \tag{7.17}$$

where $\Sigma_\varphi = \sigma_\varphi \mathbf{I}$, and

$$p_\theta(x|z_q) = \mathcal{N}(f_\theta(z_q), \sigma \mathbf{I}), \tag{7.18}$$

where $\sigma \in \mathbb{R}^+$ and $f_\theta : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^D$. Furthermore, by Bayes' theorem, the probabilistic quantization process is formulated as the Mahalanobis distance between of e_c and $z_{q,i}$ with the variance Σ_φ .

$$\hat{p}_\varphi(z_{q,i} = e_k|z, c) = \text{softmax}_k \left(\left\{ -\frac{(e_{c,j} - z_i)^T \Sigma_\varphi^{-1} (e_{c,j} - z_i)}{2} \right\}_{j=1}^K \right) \tag{7.19}$$

Here, Σ_φ is a trainable parameter that controls the degree of quantization probability during training. As σ^2 gradually decreases, Σ_φ becomes less stochastic, leading to the quantization process deterministic quantization (self-annealing).

Finally, the objective function of Categorical SQ-VAE is given by:

$$\begin{aligned}
\mathcal{L}_{csqvae} &= \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z, c)} \left[\frac{1}{2\sigma^2} \|x - p_\theta(z_q)\|_2^2 + \frac{1}{2\sigma_\varphi^2} \|z - e_c\|_2^2 \right] \\
&+ \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)} \mathcal{H}(\hat{p}_\varphi(z_q|z, c)) \\
&+ \mathbb{E}_{q_\phi(z|x)} \left[D_{KL}(q_\varphi(c|z)||p(c)) \right] + \text{const},
\end{aligned} \tag{7.20}$$

Proof of The Objective Function The ELBO of CSQ-VAE is as follows:

$$\begin{aligned} \log p_\theta(x, c) &\geq \mathcal{L}_{csqvae} = \\ \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} &\left[\log \frac{p_\theta(x)p_\varphi(z_q|x)p_\varphi(z, c|z_q)}{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z, c)} \right] \end{aligned} \quad (7.21)$$

By the Bayes' theorem, the numerator of Equation 7.21 is changed as follows:

$$\begin{aligned} p_\theta(x)p_\varphi(z_q|x)p_\varphi(z, c|z_q) \\ = p_\theta(x|z_q)p_\varphi(z_q)p_\varphi(c|z_q)p_\varphi(z|z_q) \\ = p_\theta(x|z_q)p(c)p_\varphi(z_q|c)p_\varphi(z|z_q) \end{aligned} \quad (7.22)$$

By substituting Equation 7.22 with Equation 7.21, the ELBO of CSQ-VAE is obtained as:

$$\begin{aligned} \log p_\theta(x, c) &\geq \mathcal{L}_{csqvae} = \\ \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} &\left[\log \frac{p_\theta(x|z_q)p_\varphi(z|z_q)p(z_q|c)p(c)}{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z, c)} \right] \\ = \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} &\left[\log \frac{p_\theta(x|z_q)p_\varphi(z|z_q)}{q_\phi(z|x)} \right] \\ + \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)} &\mathcal{H}(\hat{p}_\varphi(z_q|z, c)) \\ + \mathbb{E}_{q_\phi(z|x)} &\left[D_{KL}(q_\varphi(c|z)||p(c)) \right] + \text{const} \end{aligned} \quad (7.23)$$

By the Gaussian assumption, the first two terms of Equation 7.20 are obtained as:

$$\begin{aligned} \log p_\theta(x|z_q) &= \log \mathcal{N}(p_\theta(z_q), \sigma^2) \\ &= -\frac{1}{2\sigma^2} \|x - p_\theta(z_q)\|_2^2 + \frac{D}{2} \log 2\pi\sigma^2, \end{aligned} \quad (7.24)$$

where D denotes as a dimension of the data x .

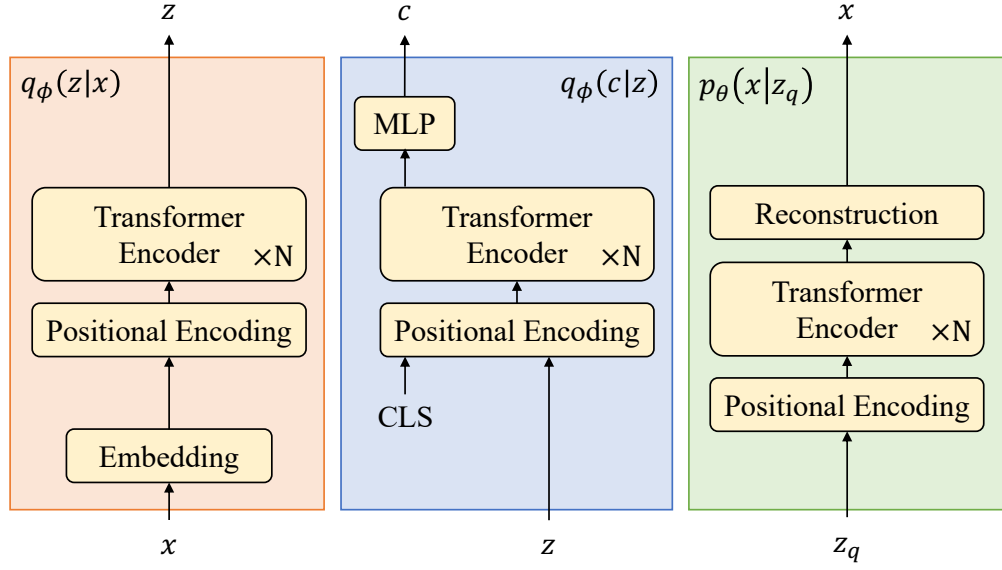


Figure 7.2: Overview of the proposed model architecture based on Figure 7.1. From left to right, the encoder $q_\phi(z|x)$, classification head $q_\phi(c|z)$, and decoder $p_\theta(x|z_q)$ are arranged.

$$\begin{aligned}
& \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} \left[\log \frac{p_\varphi(z|z_q)}{q_\phi(z|x)} \right] \\
&= \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} \left[\log \frac{\prod_{i=1}^{d_z} p_\varphi(z_i|z_q)}{\prod_{i=1}^{d_z} q_\phi(z_i|x)} \right] \\
&= \sum_{i=1}^{d_z} \mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} \left[\log \frac{p_\varphi(z_i|z_q)}{q_\phi(z_i|x)} \right] \\
&= -\frac{1}{2} \sum_{i=1}^{d_z} \left(\mathbb{E}_{q_\phi(z|x)q_\phi(c|z)p_\theta(z_q|z,c)} \left[(z_i - z_{q,i})^\top \Sigma_\varphi^{-1} (z_i - z_{q,i}) \right] \right) + \text{const}
\end{aligned} \tag{7.25}$$

7.4 Experiments

7.4.1 Model Architecture

The model architecture of Figure 7.1 is shown in Figure 7.2.

Encoder $q(z|x)$ takes x as inputs, mapping it to representations in the Embedding layer, which are then passed through the Positional Encoding and Transformer Encoder layers to produce the continuous latent variable z . The structure of Embedding layer is detailed in Table 7.1. Each block consists of a 1D Convolution layer $\rightarrow$ Group Normalization layer $\rightarrow$ SiLU [103].

Classification Head $q(c|z)$ follows a structure similar to the Vision Transformer [104]. The input z is concatenated with a learnable Classification Token and passed through the Positional Encoding layer to the Transformer Encoder. The first element of the Transformer Encoder’s output is then processed by multilayer perceptron to produce c .

Decoder $p(x|z_q)$ reconstructs x from the discrete latent variable z_q . The input z_q is passed through the Positional Encoding and Transformer Encoder layers before being converted to x in the Reconstruction layer.

The structure of Reconstruction layer is detailed in Table 7.2. The TransConv1 through TransConv3 blocks consist of a 1D Transpose Convolution layer $\rightarrow$ Group Normalization layer $\rightarrow$ SiLU, while TransConv4 consists of a 1D Transpose Convolution layer $\rightarrow$ Tanh.

Positional Encoding Each network’s Positional Encoding layer uses Rotary Encoding [105] with a learnable frequency parameter.

Model Parameters The model’s dimension was set to $d = 32$, code map size $k = 512$, with 3 Transformer Encoders per network and a Dropout Rate of 0.1. The number of attention heads in the transformer was set to 4.

7.4.2 Data Sampling

The input data consists of bounding boxes and keypoints from a human skeleton model. First, human bounding boxes are detected using YOLOv8 [95]. Then, the keypoints withing each bounding box are extracted using ViTPose [106], following the COCO keypoints format [54]. To track human movements BoostTrack+ [107] is used, generating

Table 7.1: The architecture of the Embedding layer.

Block	Kernel Size	Stride	Dimension
Conv 1	6	3	$\mathbb{R}^{15,2,90} \rightarrow \mathbb{R}^{15,d/8,29}$
Conv 2	5	3	$\mathbb{R}^{15,d/8,29} \rightarrow \mathbb{R}^{15,d/4,9}$
Conv 3	5	2	$\mathbb{R}^{15,d/4,9} \rightarrow \mathbb{R}^{15,d/2,3}$
Conv 4	3	1	$\mathbb{R}^{15,d/2,3} \rightarrow \mathbb{R}^{15,d,1}$

Table 7.2: The architecture of the Reconstruction layer.

Block	Kernel Size	Stride	Dimension
TransConv 1	3	1	$\mathbb{R}^{15,d,1} \rightarrow \mathbb{R}^{15,d/2,3}$
TransConv 2	5	2	$\mathbb{R}^{15,d/2,3} \rightarrow \mathbb{R}^{15,d/4,9}$
TransConv 3	5	3	$\mathbb{R}^{15,d/4,9} \rightarrow \mathbb{R}^{15,d/8,29}$
TransConv 4	6	3	$\mathbb{R}^{15,d/8,29} \rightarrow \mathbb{R}^{15,2,90}$

time series data for bounding boxes and keypoints. For each tracking ID, time series data is sampled in 90-frame segments (3 seconds) at intervals of 30 frames (1 seconds).

During detection, missing values may occur in human bounding boxes, requiring data cleansing and interpolation.

- **Cleansing:** time series data is discarded if $\geq 30\%$ of the frames in a 90-frame segment contain missing values.
- **Interpolation:** Linear interpolation is applied to missing values in bounding boxes and keypoints.

In operating room surveillance videos, the feet of individuals behind the operating table are often occluded. To address this, keypoints for the left knee, right knee, left ankle, and right ankle are masked. Thus, the final input data (Figure 7.1) consists of 5 face keypoints, 8 body keypoints, and the top-left and bottom-right points of the bounding box, resulting in $x \in \mathbb{R}^{90,15,2}$.

7.4.3 Scaling

The bounding boxes and keypoints are scaled using different methods. The bounding boxes are scaled on the video frame, with the center at (0,0), the top-left corner at (-1,-1), and the bottom-right corner at (1,1). Therefore, the bounding boxes represent spatial movement across the entire frame. The keypoints are scaled within each bounding box, with the center at (0,0), the top-left corner at (-1,-1), and the bottom-right corner at (1,1). Therefore, the keypoints represent detailed body movements, such as orientation and posture.

7.4.4 Role Labels

The role labels are defined as follows: surgeon, anesthetist, scrub nurse, circulator, and other. "Other" includes observers who are not directly involved in surgeries, such as trainees

7.4.5 Annotation Protocol

We describe the annotation protocol for semi-supervised learning. To ensure reproducibility, we conducted annotations and selected a subset of the dataset as follows:

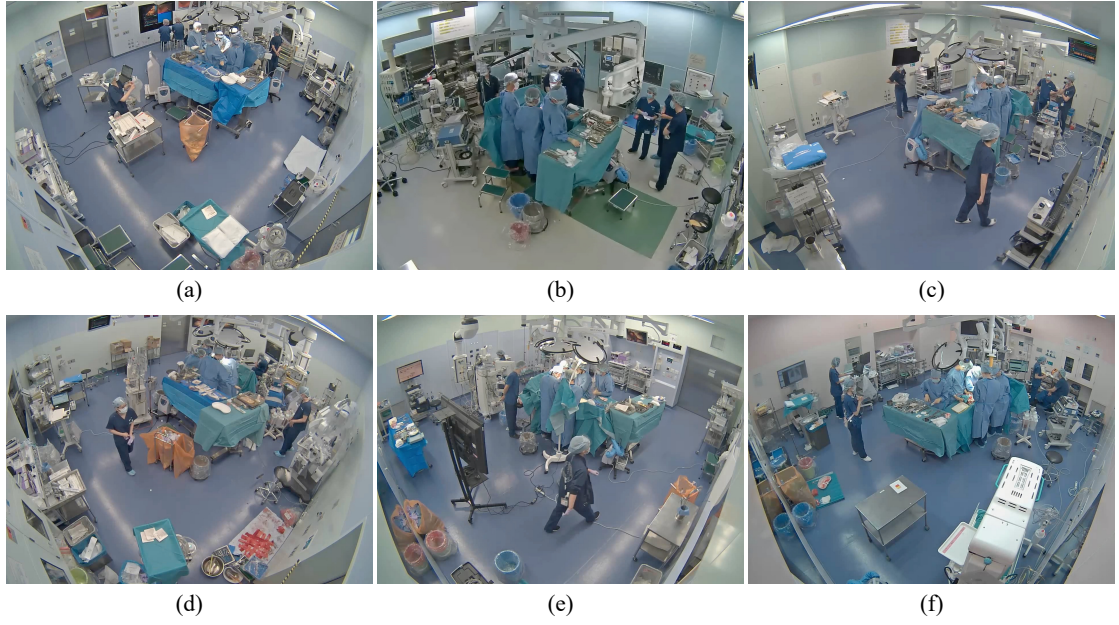


Figure 7.3: Sample frames of operating room surveillance videos. The frames from Surgery 1 to Surgery 6 are arranged in order from (a) to (f).

1. Cleansing

- Tracking IDs where changes have occurred.
- Individuals with different positions during surgery (e.g., surgeons entering and exiting for shifts, anesthesiologists, scrub nurses).

2. Label each tracking ID.

3. Shuffle the tracking IDs (using seed 42).

4. Select tracking IDs until the number of samples exceeds 1000.

Table 7.3: The number of samples of training dataset.

	Surgeon	Anesthetist	Scrub Nurse	Circulator	Other	Non-Labeled	Total
S1	1,315	1,079	1,015	1,021	1,001	14,119	19,550
S2	2,212	1,103	1,149	1,008	1,068	17,956	24,496
S3	1,056	1,130	1,023	1,210	1,235	13,583	19,237

Table 7.4: The number of samples of test dataset.

	Surgeon	Anesthetist	Scrub Nurse	Circulator	Other	Total
S1	2,430	1,096	1,707	1,306	3,581	10,120
S2	3,663	1,028	1,607	1,915	4,701	12,914
S3	1,824	2,800	1,641	1,594	2,609	10,468
S4	698	1,138	1,520	1,731	331	5,418
S5	4,071	1,896	1,797	620	82	8,466
S6	4,544	3,186	1,798	3,270	15	12,813

7.4.6 Contents of Dataset

The datasets used in our experiments were consisted of various samples from operating room surveillance video. As shown in Figure 7.3, the data includes different scenarios captured in the operating room. Frames (a) to (c) depict samples utilized in both the training and test datasets. These images include scenes involving the surgeon, anesthesiologist, scrub nurse, circulator, and other personnel, providing a comprehensive dataset for training the model. Frames (d) to (f), on the other hand, were exclusively used in the test dataset. These samples were chosen to evaluate the model’s performance in scenarios not encountered during training.

Training Dataset The training data was created by extracting 1 hour videos from operating room surveillance videos of three different surgeries using the Data Sampling and Annotation Protocol described above. Each surgery took place in a different operating room. The composition of the dataset is shown in Table 7.3.

Test Dataset The test data was created by extracting 30 minutes videos from surveillance camera footage of six operating rooms using the Data Sampling and up to step 2 of the Annotation Protocol described above. Three of these videos were created from films of surgeries included in the training dataset, and the other three from operating rooms not used for training. The composition of the dataset is shown in Table 7.4.

7.4.7 Training Algorithm

Table 7.5 shows the training conditions. Due to the varying arrangements of medical instruments as seen from the surveillance cameras in different operating rooms, we trained the model using two patterns to evaluate these effects.

Pattern 1: One model was trained using the entire dataset, as indicated in T1.

Pattern 2: Other models were trained using the S1, S2, and S3 datasets, as indicated in T2, T3, and T4.

To approximately equalize the number of training steps, Pattern 1 was trained for 100 epochs, while Pattern 2 was trained for 300 epochs. Across all training conditions, the optimizer was AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, the learning rate was set to 0.005 and the batch size was set to 128.

7.4.8 Results

Training Dataset As shown in Table 7.6, under learning condition T1, S1 and S3 achieved an accuracy exceeding 0.96, while S2 reached 0.774. Conversely, under learning conditions T2 through T4, accuracy improved across all datasets compared to T1. S1 and S3 exceeded 0.98 in accuracy, and S2 achieved 0.91.

Figure 7.4 shows the confusion matrices for the training dataset. Under learning condition T1, only other in S2 was misclassified as circulator, resulting in a recognition accuracy below 0.5. However, under all other learning conditions and across all datasets, the accuracy of each role label exceeded 0.84. In particular, for S1 and S3, accuracy for all labels in all learning conditions surpassed 0.9.

Test Dataset As shown in Table 7.7, the test datasets S1 to S3 used in training achieved their highest accuracies under learning conditions T2 to T4, with maximum values of 0.973, 0.854, and 0.958, respectively. Similarly, S4, like S1, had its highest accuracy under learning condition T2, with a maximum value of 0.891. In contrast, S5 and S6 had lower accuracies, with maximum values of 0.254 and 0.642, respectively.

Table 7.5: The training conditions. The checkmarks indicate which datasets were used for each training condition.

Name	Pattern	S1	S2	S3
T1	1	✓	✓	✓
T2	2	✓		
T3	2		✓	
T4	2			✓

Table 7.6: The accuracy scores of training dataset.

Name	Pattern	S1	S2	S3
T1	1	0.964	0.774	0.968
T2	2	0.985	–	–
T3	2	–	0.910	–
T4	2	–	–	0.988

Table 7.7: The accuracy scores of test dataset.

Name	Pattern	S1	S2	S3	S4	S5	S6
T1	1	0.966	0.718	0.866	0.854	0.254	0.577
T2	2	0.973	0.126	0.201	0.891	0.161	0.559
T3	2	0.050	0.854	0.051	0.193	0.074	0.052
T4	2	0.576	0.100	0.958	0.496	0.160	0.642

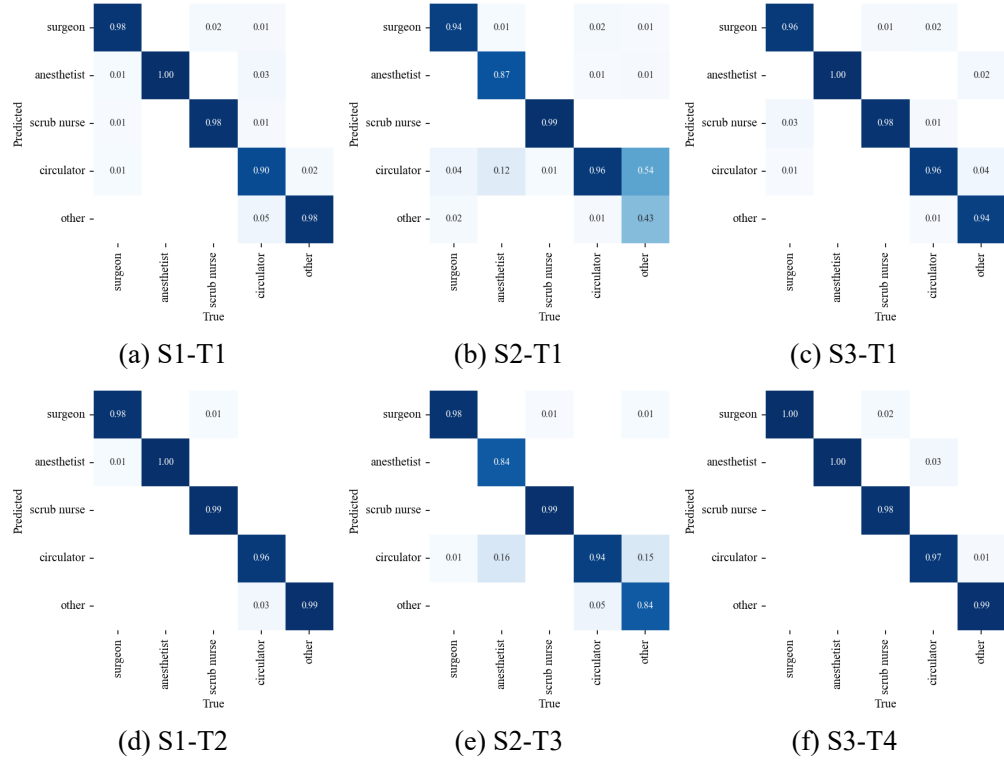


Figure 7.4: Confusion matrices of training dataset. The results for S1 to S3 under training condition T1 are arranged in order from (a) to (c). The result for S1 under condition T2 in (d), for S2 under condition T3 in (e), and for S3 under condition T4 in (f).

As shown in Figure 7.5, under learning condition T1, accuracies for S1 to S4 were high. Specifically, the surgeon, anesthetist, and scrub nurse were all classified with accuracies exceeding 0.9. However, the circulator and other roles were sometimes confused, with minimum accuracies of 0.78 and 0.24, respectively. While some roles were correctly classified in S5 and S6, such as the scrub nurse in S5 and the anesthetist in S6, most classifications had low accuracy.

As shown in Figure 7.6, under learning condition T2, the accuracies for S1 and S4 were high. Specifically, the surgeon, anesthetist, scrub nurse, and circulator were all classified with accuracies exceeding 0.8. However, other in S4 was sometimes confused with the circulator, resulting in an accuracy of 0.30. In S2 and S3 were classifications were circulators. In S5 and S6, only the circulator had a high accuracy exceeding 0.9, while accuracies for other roles were low.

As shown in Figure 7.7, under learning condition T3, only S2 achieved high accuracy. All roles except other were classified with accuracies exceeding 0.9. The other role was sometimes confused with the circulator, resulting in an accuracy of 0.65. In datasets other than S2, most classifications were frequently assigned to the circulator.

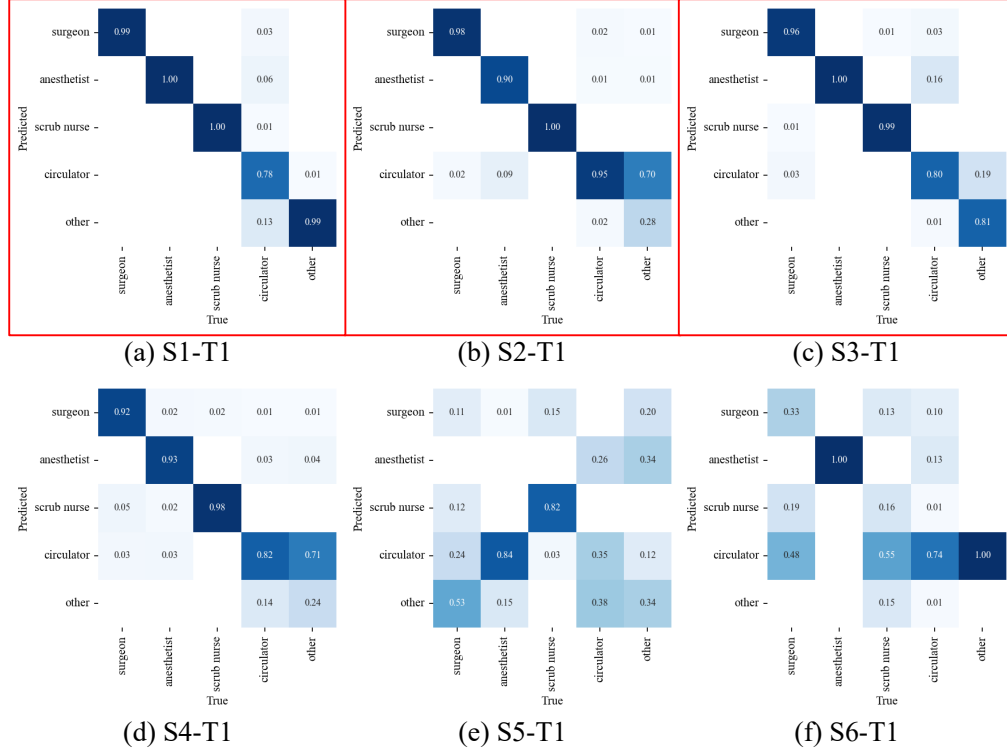


Figure 7.5: Confusion matrices of the test dataset. The results from S1 to S6 under training condition T1 are arranged in order from (a) to (f). The red-framed results were used for training.

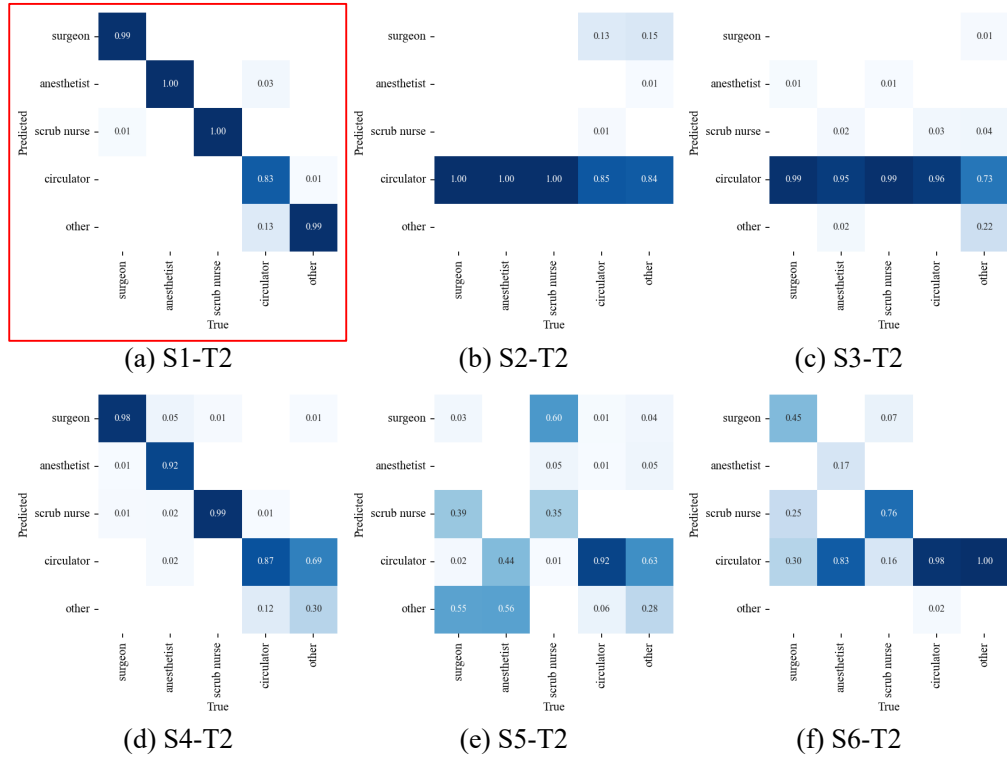


Figure 7.6: Confusion matrices of the test dataset. The results from S1 to S6 under training condition T2 are arranged in order from (a) to (f). The red-framed result was used for training.

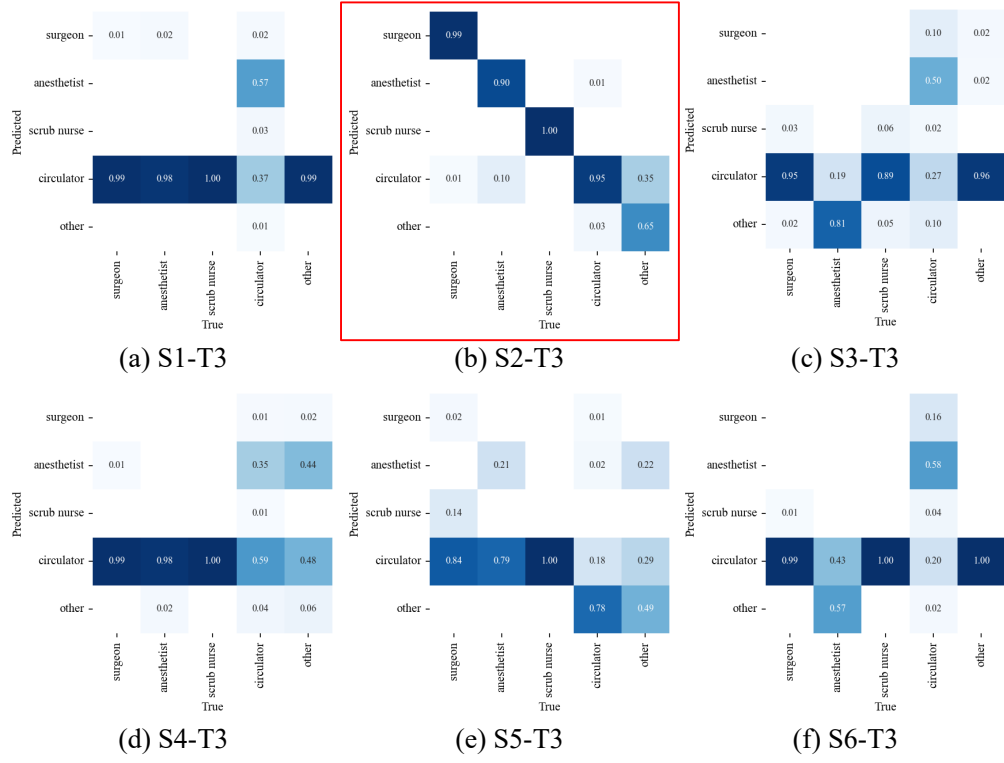


Figure 7.7: Confusion matrices of the test dataset. The results from S1 to S6 under training condition T3 are arranged in order from (a) to (f). The red-framed result was used for training.

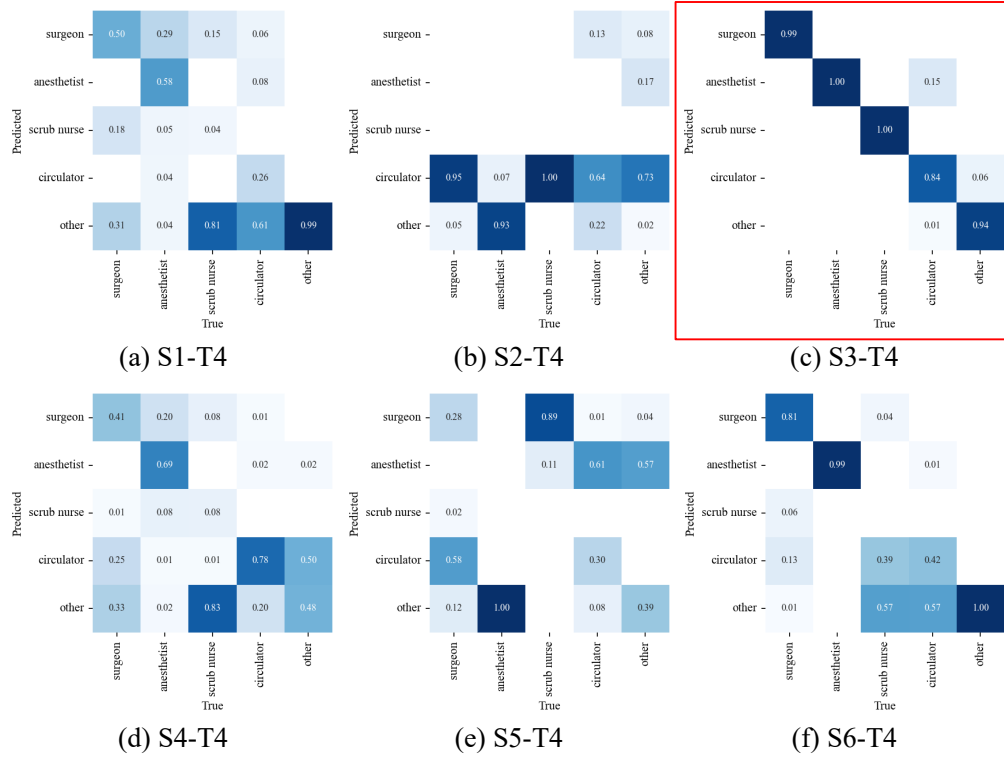


Figure 7.8: Confusion matrices of the test dataset. The results from S1 to S6 under training condition T4 are arranged in order from (a) to (f). The red-framed result was used for training

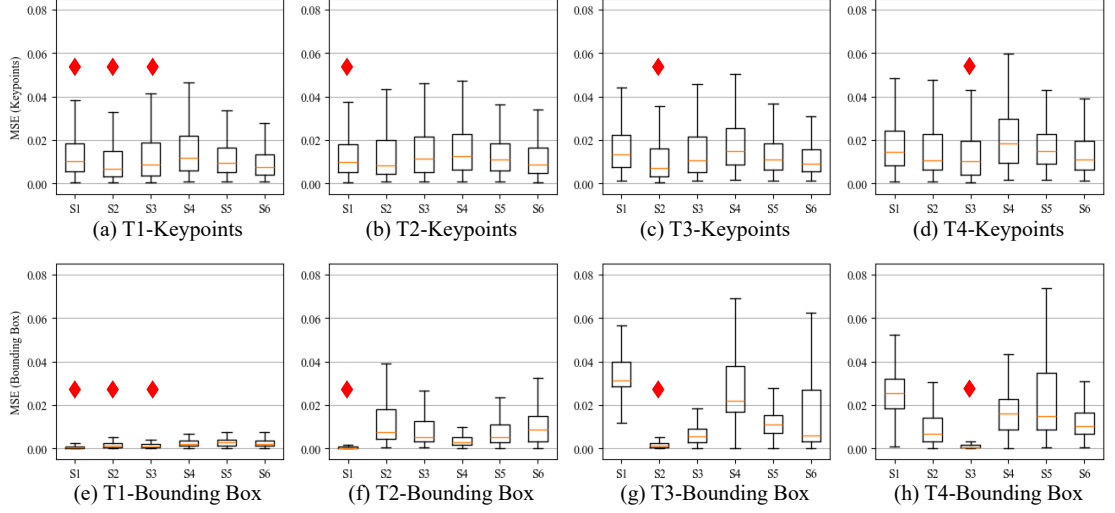


Figure 7.9: Box plots of MSE for keypoints and bounding boxes under each learning condition. (a) to (d) show the results for keypoints, and (e) to (h) show the results for bounding boxes. The red diamonds indicate that the surgeries were used for training.

As shown in Figure 7.8, under learning condition T4, only S3 achieved high accuracy. All roles were classified with accuracies exceeding 0.84. In S1, S4, and S6, some labels were classified with high accuracy. For example, other in S1 and S6 had an accuracy of approximately 1.0, and the circulator in S4 had an accuracy of 0.78. In S2 and S5, no roles were classified with high accuracy.

Figure 7.9 shows the box plot of the MSE for the input data and reconstruction results of keypoints and bounding boxes. As shown in Figure 7.9 (a) to (d), there was no clear trend in the MSE of keypoints. Across all learning conditions, no discernible patterns were observed in the reconstruction results between the datasets used for training and those not used for training. In contrast, as shown in Figure 7.9 (e) to (h), the MSE for bounding boxes was extremely low for the datasets used in training. Furthermore, under learning condition T1, the MSE remained low even for datasets S4 to S6 which were not used for training.

7.5 Discussion

We analyzed the spatial distribution of misclassified predictions to examine the characteristics of the model’s role classification. Figure 7.10 and Figure 7.11 show the distribution of bounding boxes, colored by ground truth and false predictions, for the model trained under learning condition T1.

First, we consider S1 through S4, which achieved high classification accuracy. As shown in Figure 7.5, these surgeries had classification accuracies exceeding 0.9 for roles

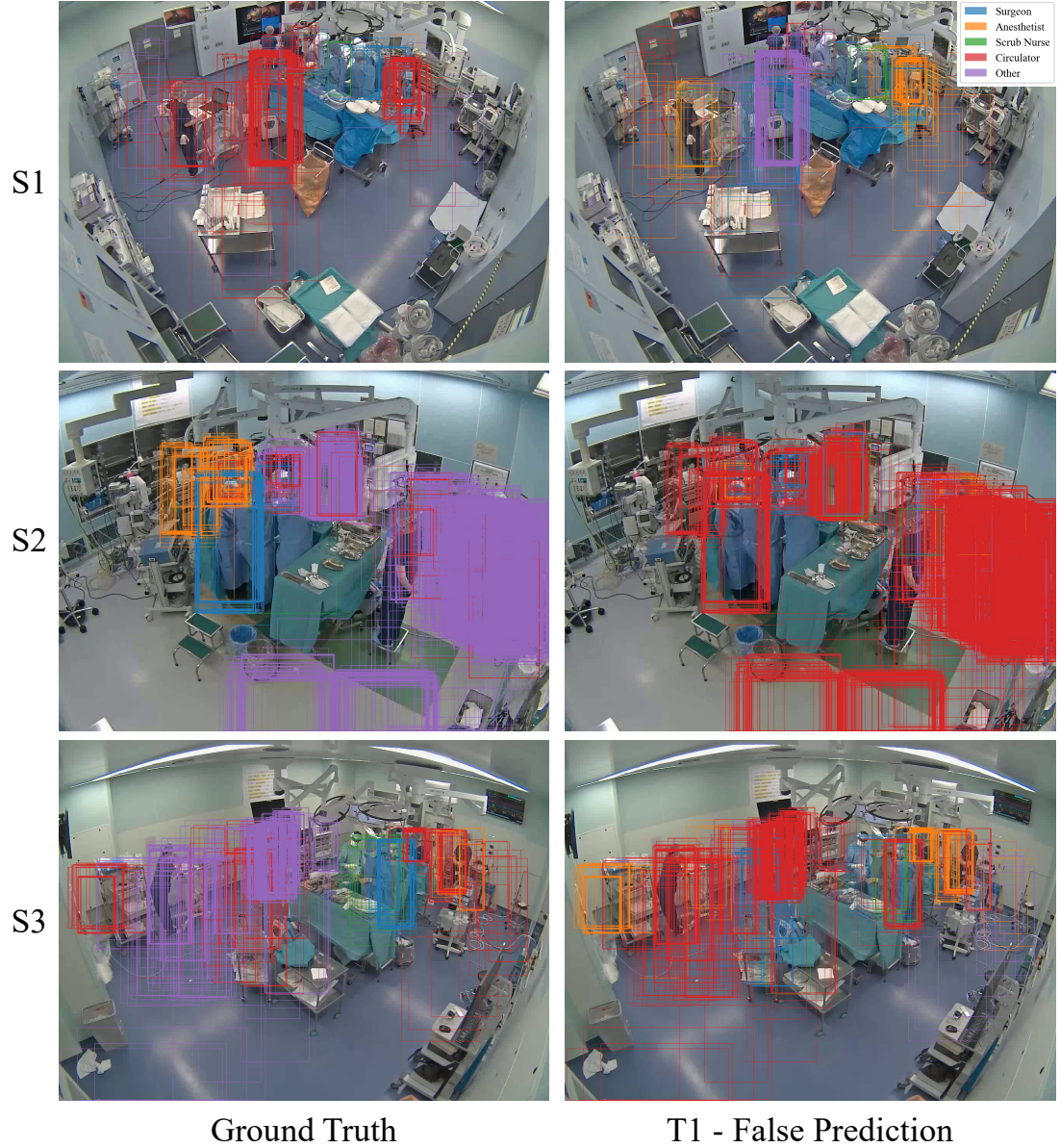


Figure 7.10: Scatter plot of bounding boxes for misclassified samples from S1 to S3. The left column shows bounding boxes colored by the ground truth, while the right column shows bounding boxes colored by the falsely predicted results under training condition T1.

such as surgeon, anesthetist, and scrub nurse, whose positions remained relatively stable. In contrast, accuracy was lower for roles of circulators and others, who moved frequently and had variable positions.

As shown in Figure 7.5 (a), (c), and (d), the classification accuracy for circulators in S1, S3, and S4 was lower than for other roles. In Figure 7.10 for S1, circulators were misclassified as others near the center of the video frame. Similarly, in Figure 7.11 for S4, circulators were also misclassified as others near the center. As shown in Figure 7.12,

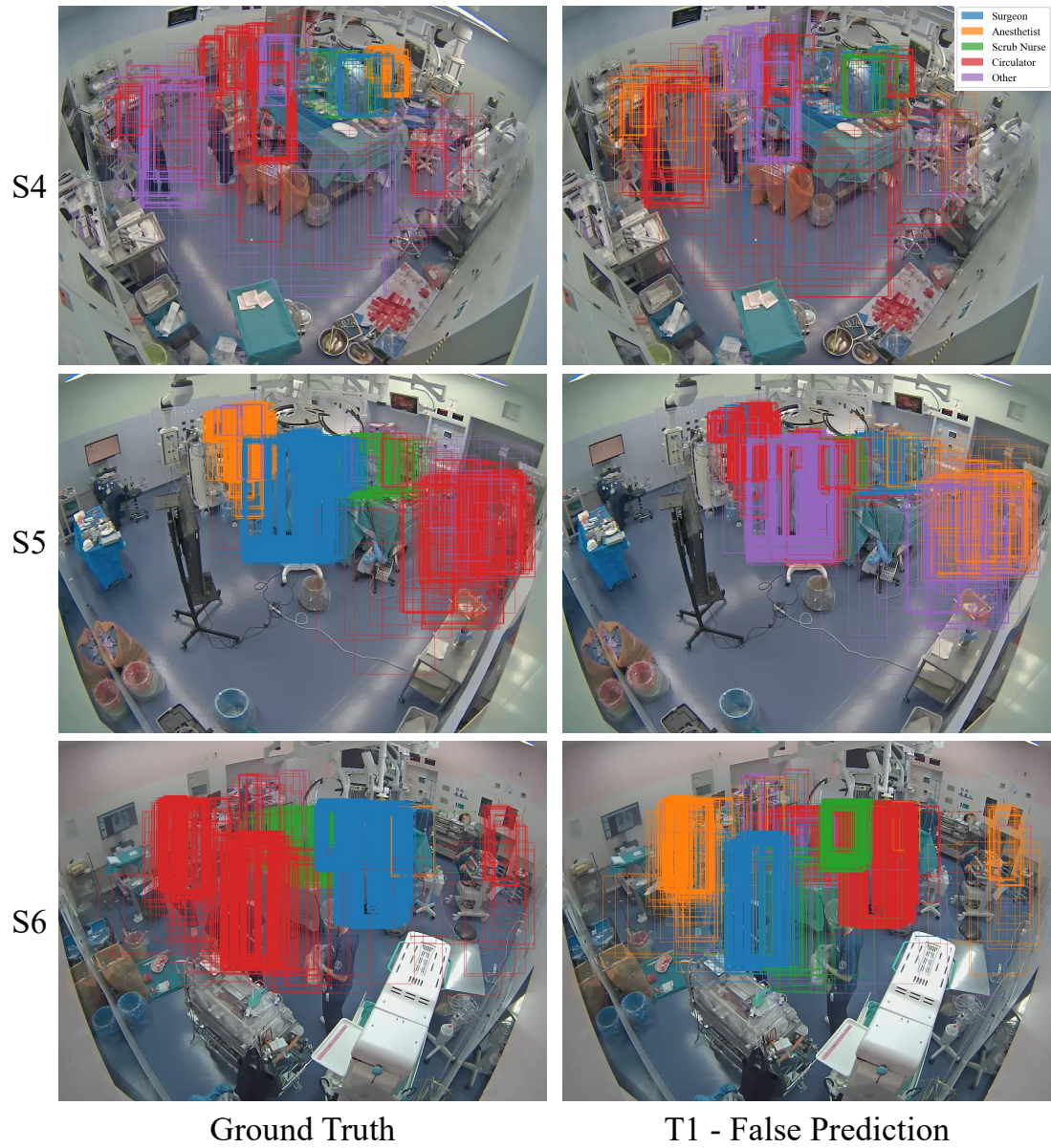


Figure 7.11: Scatter plot of bounding boxes for misclassified samples from S4 to S6. The left column shows bounding boxes colored by the ground truth, while the right column shows bounding boxes colored by the falsely predicted results under training condition T1.

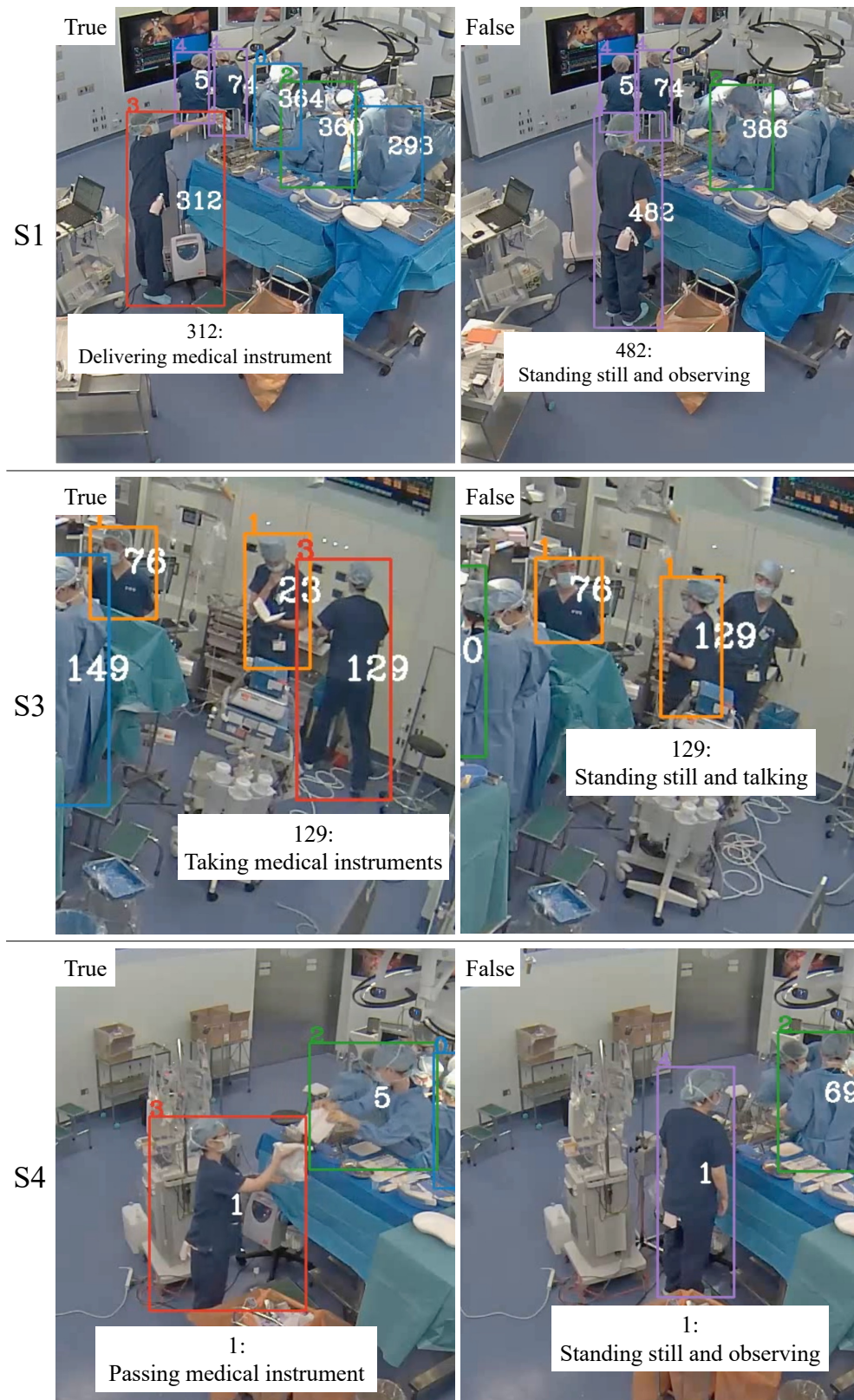


Figure 7.12: Correctly classified samples (left) and misclassified samples (right) for circulator (label 3).

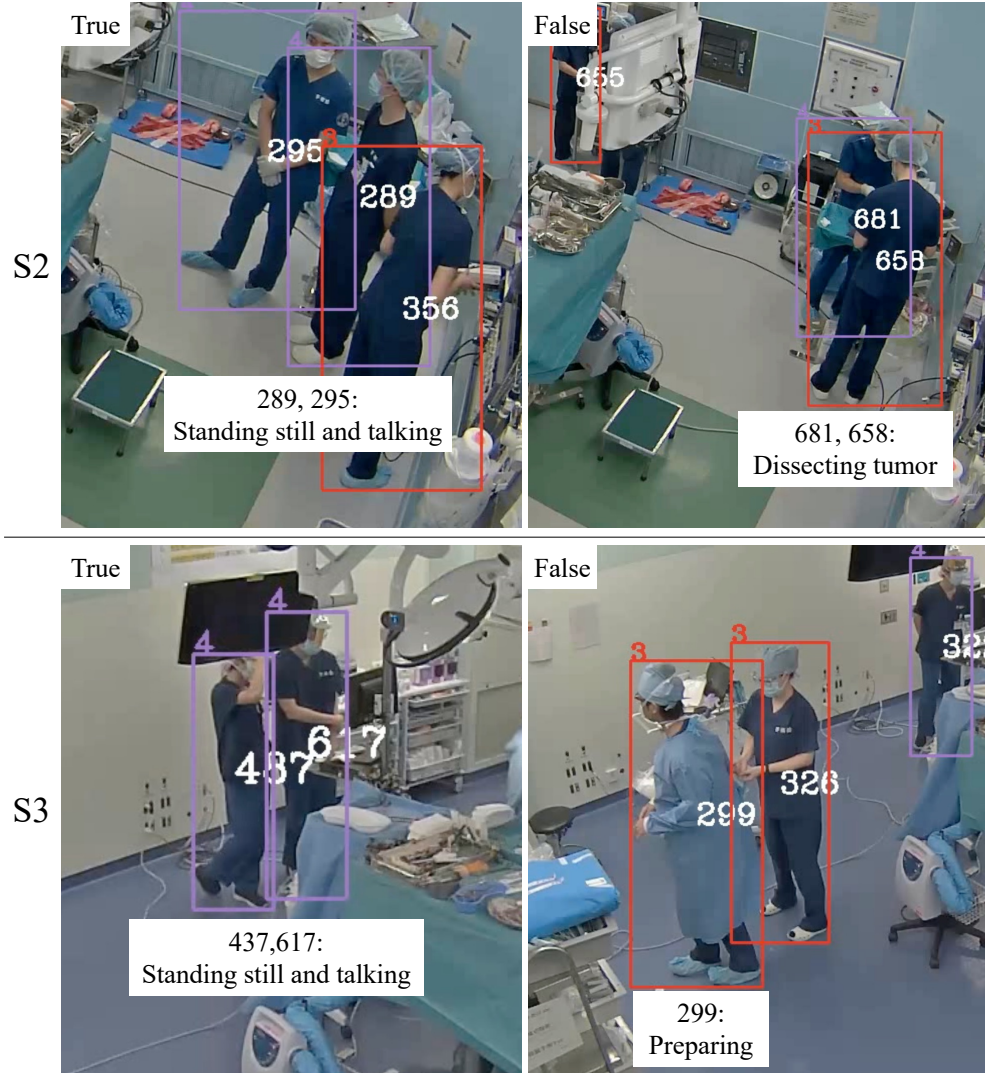


Figure 7.13: Correctly classified samples (left) misclassified samples (right) of other (label 4).

circulators in S1 and S4 were correctly classified when delivering medical instruments but misclassified as others when observing the surgical field from the pedestal. In Figure 7.10 for S3, circulators were misclassified as anesthetists on the right side of the video frame. As shown in Figure 7.12 for S3, circulators conversing with anesthetists near the anesthesia machine were misclassified. These results suggest that the model classifies roles based on position when stationary, identifies circulators based on significant spatial movement, and considers the magnitude of movement.

Furthermore, as shown in Figure 7.5 (b), (c), and (d), classification accuracy for others in S2, S3, and S4 was lower than for other roles. In Figure 7.10 for S2, individuals labeled as others were misclassified as circulators on the right side of the video frame. As shown in Figure 7.13 for S2, individuals facing the wall during tumor dissection were classified as circulators. In Figure 7.10 for S3, individuals categorized as others were

misclassified as circulators on the left side of the video frame. As shown in Figure 7.13 for S3, similar to S2, surgical team members preparing for shift changes (labeled as others since they were not directly involved in surgery) were classified as circulators due to facing the wall. These findings suggest that the model considers body orientation for role classification. However, in Figure 7.11 for S4, others were misclassified as circulators on the left side of the video frame, despite no samples labeled as others in this position in the training dataset from S1 to S3.

In Figure 7.11 for S5, surgeons and anesthetists were misclassified as circulators on the left side of the frame, while circulators were misclassified as anesthetists and others on the right. As discussed in Section 7.4.6, S5 has a similar configuration to S2, but its role classification predictions resembled those of S1 and S3. Comparing Figure 7.3 (b) and (e), the operating table in S2 (b) is closer to the surveillance camera, whereas in S5 (e), its position is similar to S3 (c) but with a different orientation. In Figure 7.11 for S6, circulators were misclassified as surgeons, anesthetists, and others on the left side of the frame, while surgeons were misclassified as circulators on the right. The misclassified circulators occupied positions similar to those of others in S1, while the misclassified surgeons and anesthetists occupied positions similar to those in S2. The surgeon's position was similar to that of a circulator observing the surgical field in S2.

Since the input data's spatial features were limited to bounding boxes, role classification errors likely occurred in positions not present in the training data. Incorporating relative spatial features, such as distance and direction from the operating table, could improve classification performance for unseen data.

7.6 Conclusion

In this chapter, we analyzed the activities of surgical team members captured in operating room surveillance videos to identify their specific roles. To achieve this, we introduced a novel role classification model using semi-supervised learning. Our model, which employs a categorical quantized variational autoencoder, enables role classification of surgical team members by maintaining distinct latent variable spaces for each role. This allows for end-to-end training while capturing activity features specific to each role.

Surgeries included in the training dataset achieved role classification accuracies of 0.75 or higher in the test dataset. Surgeries not included in training also achieved classification accuracies of 0.85 or higher, depending on the operating table configuration in

the surveillance video frames. Additionally, our results suggest that the model classifies roles by considering activity features such as the magnitude of spatial movement.

Part III

Intraoperative Situational Analysis

Chapter 8

Intraoperative Situation Recognition and Event Detection

8.1 Introduction

In this chapter, we discuss situation recognition and event detection during surgery by applying the activity recognition models proposed in Part I and Part II.

In Part I, we recognized specific group activities during surgery using individual activity features extracted from key points in the skeleton model. Specifically, we detected Medical Instrument Passing (MIP) and identified objects or regions of interest through Group Attention (GA). Group activity recognition was applied to operating room surveillance videos, and the results were analyzed to assess their potential for assisting in event detection and situation recognition.

In Part II, we proposed a model that performs end-to-end individual activity recognition using deep learning. The model proposed in Chapter 7 achieved high accuracy in classifying roles and reconstructing the input activities. Based on role classification results, we attempted event detection and situation awareness by analyzing temporal changes in the number of individuals by role and variations in internal feature.

8.2 Linear Transformation-Based Group Activity Recognition

8.2.1 Methods

Medical Instrument Passing

The Medical Instrument Passing (MIP) model proposed in Chapter 3 is applied to operating room surveillance videos. The weights learned in the experiments in Section 3.3.2 were used to perform MIP recognition.

Group Attention

The method proposed in Chapter 4 Group Attention (GA) is applied to the entire operating room surveillance videos. The maximum value of the Attention Probability (AP) calculated by Equation 4.2 was computed for the entire operating room, and its temporal transition was measured. The peaks and troughs were detected using the following algorithm.

The Group Attention (GA) method proposed in Chapter 4 is applied to operating room surveillance videos. The maximum value of the Attention Probability (AP), calculated using Equation 4.2, was computed for the entire operating room, and its temporal transitions were analyzed. Peaks and troughs were detected using the following algorithm:

1. The maximum value of AP is smoothed using a moving average over 1800 frames.
2. Peaks and troughs are detected by setting the prominence threshold to 0.3.
3. Among the detected peaks, only those with AP values higher than 1.5 are extracted.
4. Among the detected troughs, only those with AP values higher than 0.5 are extracted.

8.2.2 Dataset

Table 8.1 shows the contents of the dataset to which handcrafted feature-based group activity recognition was applied. All of dataset are operating room surveillance videos that recorded all scenes from the preparation for surgery to the cleanup. Check marks were added to those used in both Chapter 3 and Chapter 4 of Part I.

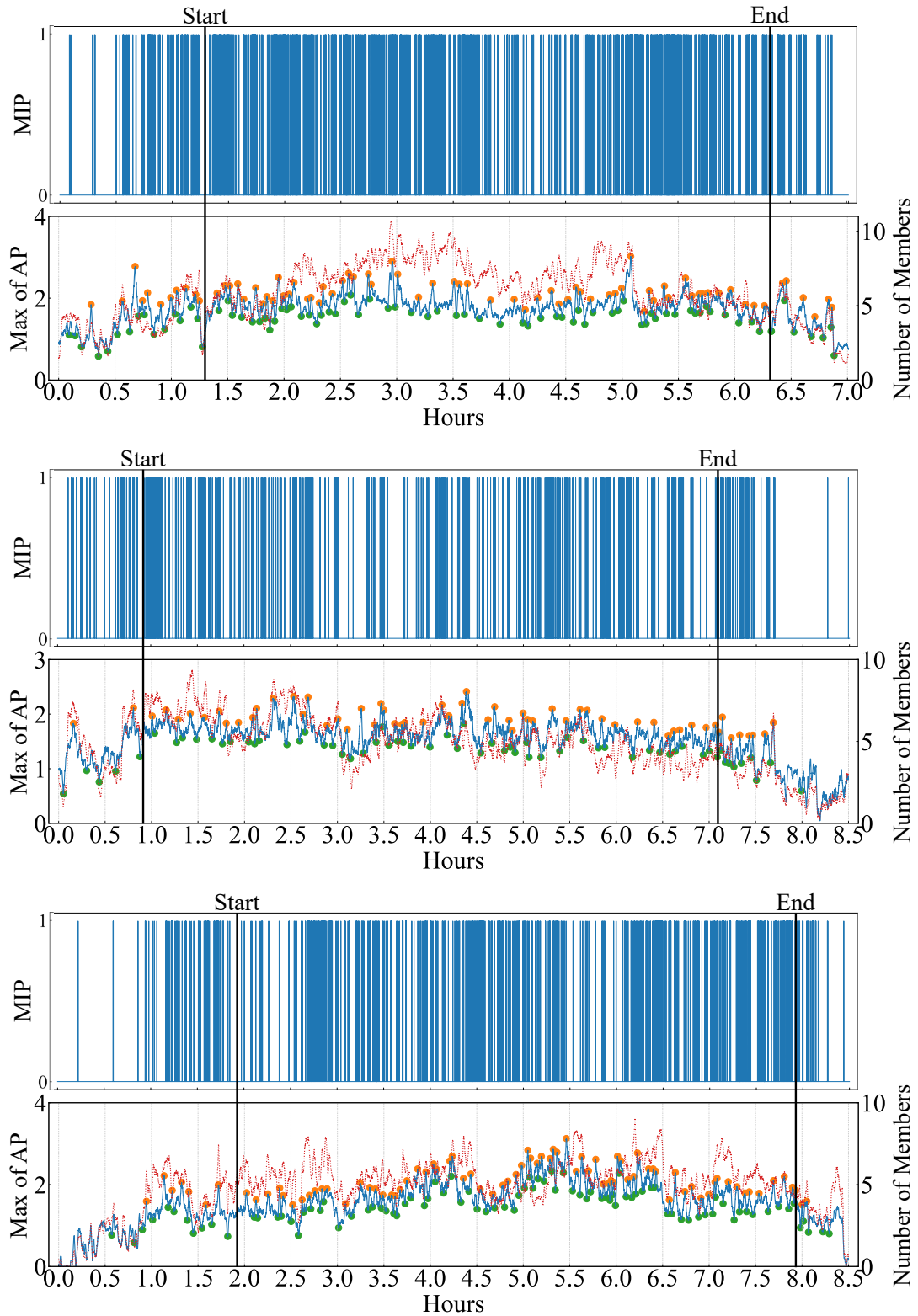


Figure 8.1: Results of group activity recognition for Surgery 1 (top), Surgery 2 (middle), and Surgery 3 (bottom).

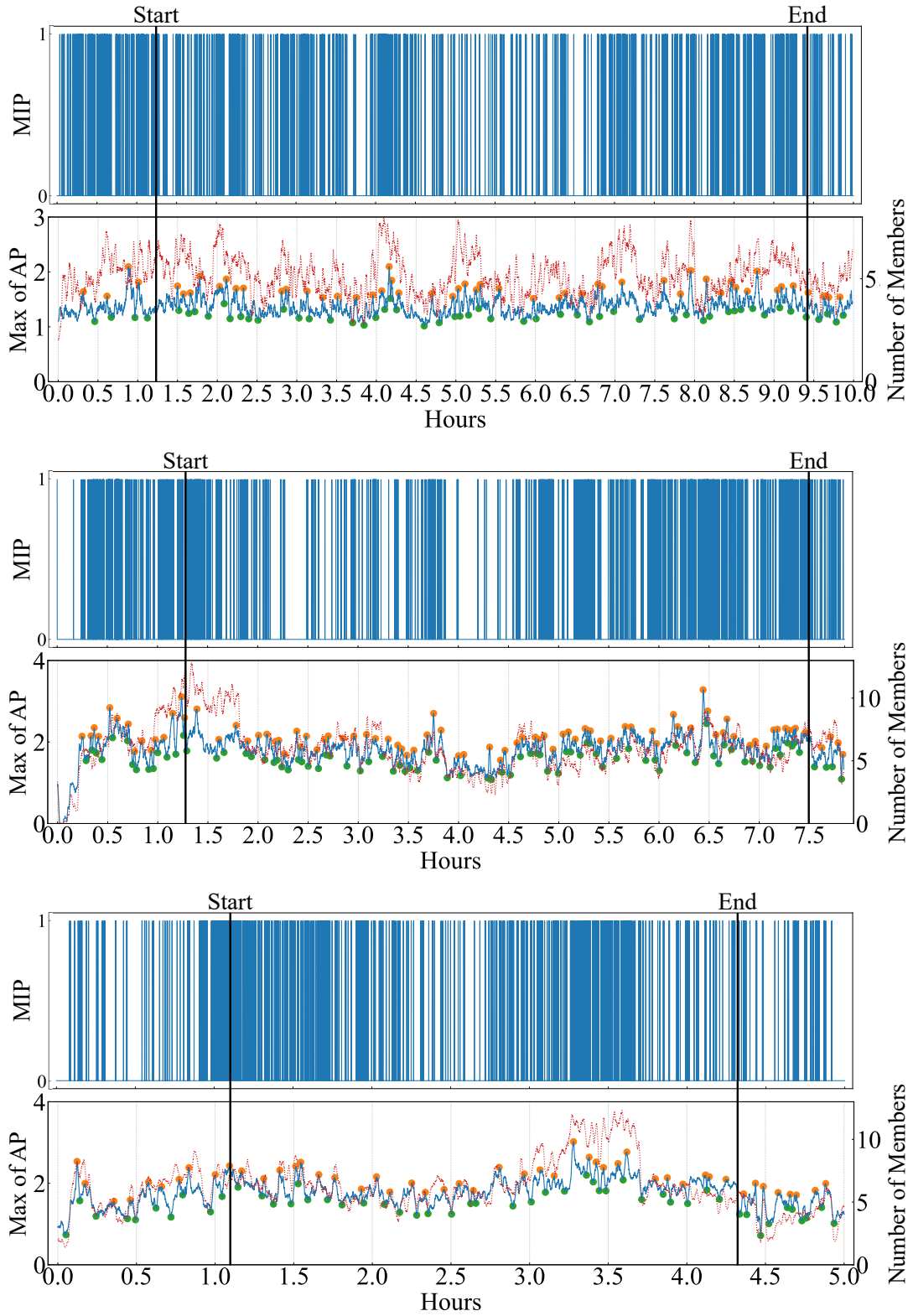


Figure 8.2: Results of group activity recognition for Surgery 4 (Top), Surgery 5 (Middle), and Surgery 6 (Bottom).

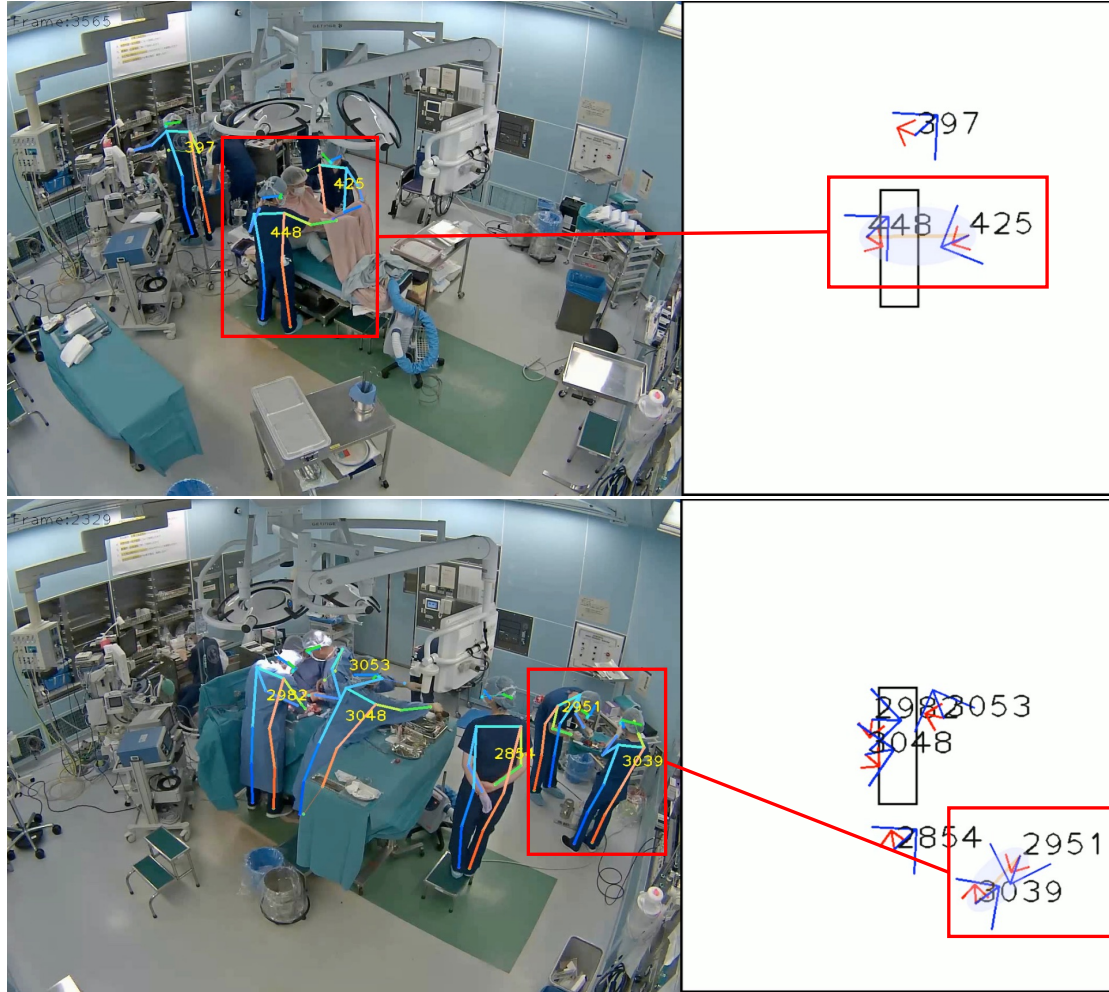


Figure 8.3: Examples of activities recognized by the MIP model involving two surgical team members working together. (Top)Two surgical team members cover a patient with a blanket. (Bottom)Two surgical team members dissect the excised tumor.

8.2.3 Results

Figure 8.1 and Figure 8.2 show the recognition results of MIP and the transition of the maximum AP value calculated from GA. The upper part of each result indicates whether MIP was detected, with 1 representing detection and 0 indicating no detection.

Table 8.1: Contents of operating room surveillance videos used for event detection and situation recognition with handcrafted feature-based group activity recognition.

	Time Length	Included in Part I
Surgery 1	07:00:32	✓
Surgery 2	08:30:00	
Surgery 3	08:29:02	✓
Surgery 4	09:29:07	
Surgery 5	07:50:36	✓
Surgery 6	05:00:00	

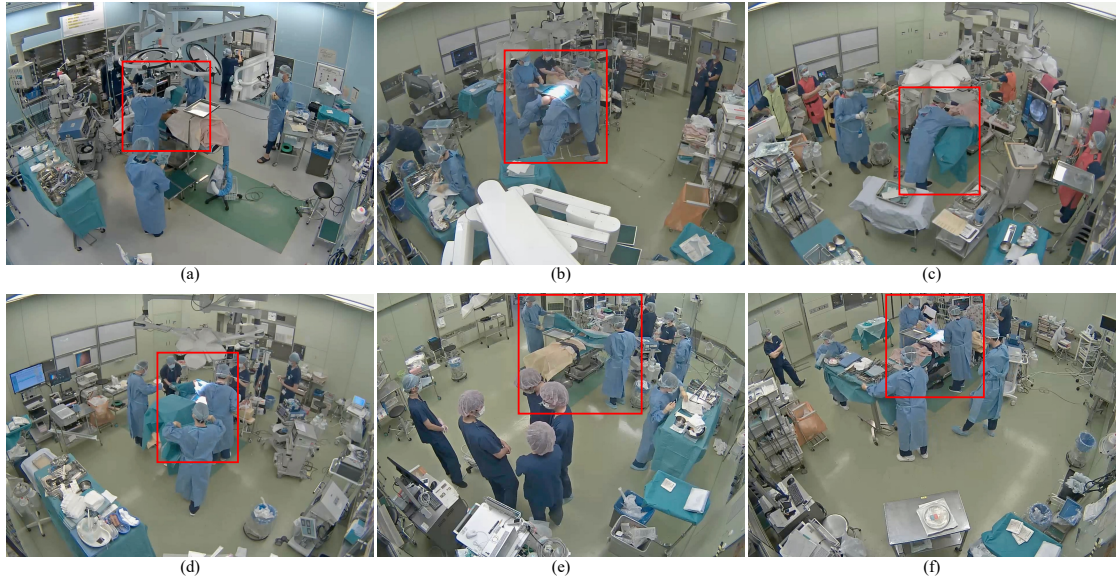


Figure 8.4: Frames showing the draping process in each operating room surveillance video (“Start” in Figure 8.1 and Figure 8.2). The surgeons performing the draping are highlighted with red boxes. Results from surgeries 1 to 6 are arranged in order from (a) to (f).

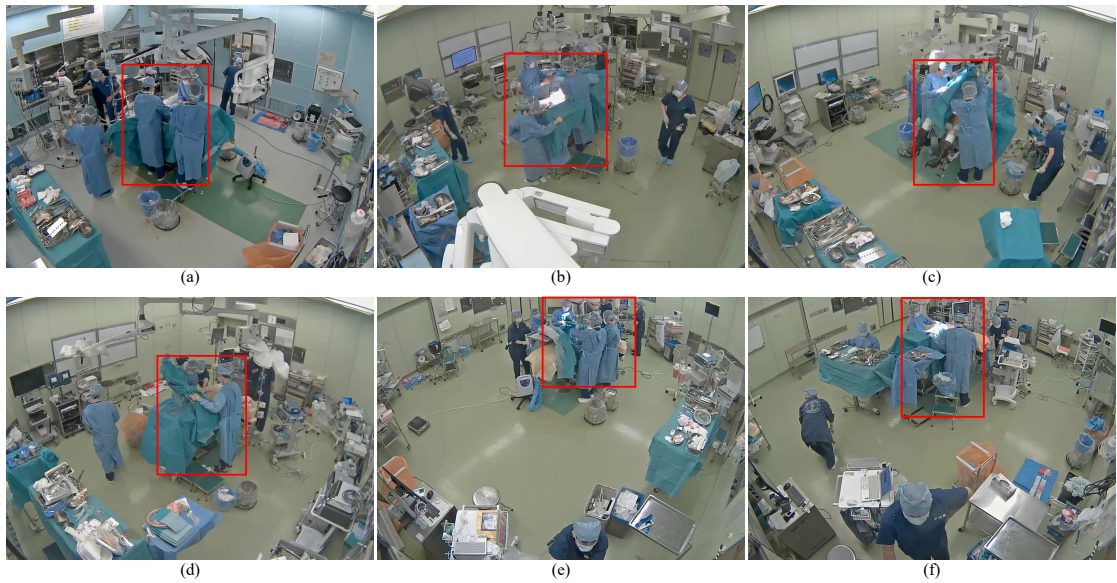


Figure 8.5: Frames showing the draping removal process in each operating room surveillance video (“End” in Figure 8.1 and Figure 8.2). The surgeons removing the draping are highlighted with red boxes. Results from surgeries 1 to 6 are arranged in order from (a) to (f).

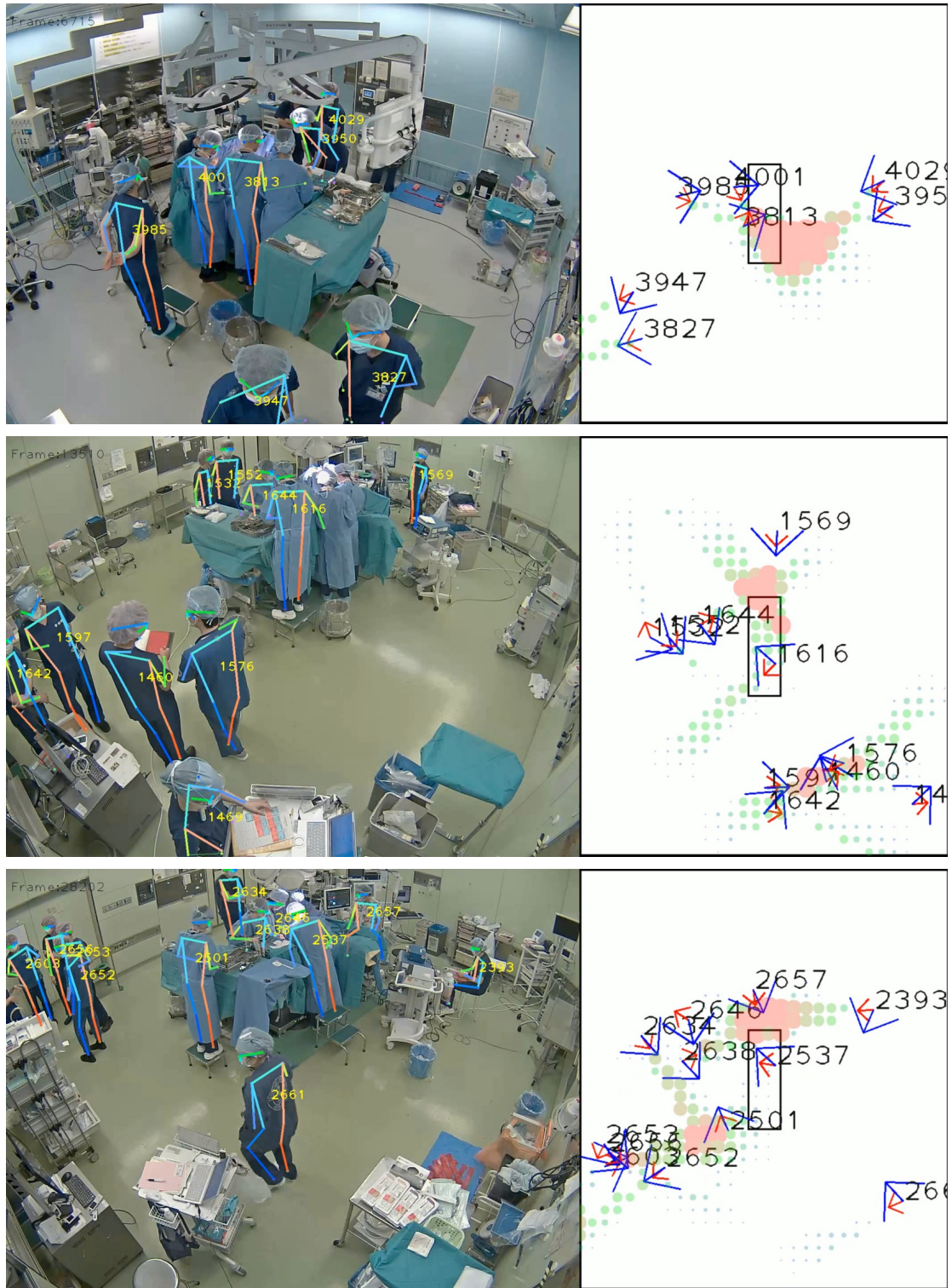


Figure 8.6: Examples of discrepancies between AP movements and the number of surgical team members. (Top) Results around the 3 hours mark of Surgery 1. Interns (ID:3827 and 3947) are conversing in front of the monitor. (Middle) Results around the 3 hours 7 minutes mark of Surgery 5. Interns (ID:1460, 1576, 1597, and 1642) are conversing in a corner of the operating room. (Bottom) Results around the 3 hours 15 minutes mark of Surgery 6. Interns (ID:2603, 2652, 2653, and 2656) are conversing in a corner of the operating room.

The lower part of each result displays the transition of the maximum AP value within the operating room (blue solid line, left axis) and the transition of the number of people detected within the operating room (red dotted line, right axis). Orange dots represent peaks in the transition of the maximum AP value, while green dots indicate troughs. Additionally, "Start" marks the timing of draping, and "End" marks the timing of draping removal.

Medical Instrument Passing

MIP recognition was performed on operating room surveillance videos that were not included in the training dataset of Section 3.3. As shown in Figure 8.3, an activity was detected in which two surgical team members faced each other and moved a common object together. As described in Section 3.3.2.3, the model was trained to prioritize recall, leading to recognition of activities beyond MIP. Consequently, the model is capable of recognizing collaborative activities involving two individuals, in addition to MIP. Note that the upper results of Figure 8.1 and Figure 8.2 also include the timings of activities other than MIP.

Additionally, at the "Start" and "End" timings, covering and removing the draping were performed by two surgeons, respectively. We expected the MIP model to detect this collaboration. Figure 8.4 shows the draping scenes in each operating room. Except for Surgery 3 (c), draping was performed by two members, whereas Surgery 3, it was performed by a single member. Examining Figure 8.1 and Figure 8.2, we confirmed that the detection frequency of MIP in Surgery 3 was lower than in other surgeries. Additionally, Figure 8.5 shows the draping removal process in each operating room. It was confirmed that two surgeons cooperated in removing the draping in all surgeries. This corresponded with a high frequency of MIP detection, as seen in Figure 8.1 and Figure 8.2. Surgery 2 showed an increased frequency of MIP detection after the draping removal timing. These findings suggest that when adapted to multi-class classification, the MIP model could effectively recognize two-person collaborations, providing valuable insights for understanding surgical situations and event detection.

Group Attention

The AP calculated by Equation 4.2 is proportional to the number of surgical team members. Therefore, as shown in Figure 8.1 and Figure 8.2, the maximum value of the AP (blue solid line) and the number of surgical team members (red dashed line) showed similar trends over the long term. Additionally, for most of the duration of all surgeries, the maximum AP values were concentrated near the operating table. However, there

were periods where only the number of surgical team members significantly increased: between 2 hours and 5 hours in Surgery 1, between 1 hours and 2 hours in Surgery 5, and between 3 hours and 3.7 hours in Surgery 6. As shown in Figure 8.6, this occurred because interns entered the room and engaged in conversations in areas away from the operating table. By identifying discrepancies between the trends of AP and the number of surgical team members, it is possible to detect groups that are focused on areas other than the surgical field during surgery. Additionally, the presence of a non-intern group could indicate that the occurrence of a specific event.

8.3 End-to-End Individual Activity Recognition

8.3.1 Methods

We applied the role classification model, proposed and trained in Chapter 7 to the entire set of operating room surveillance videos. Our goal is to achieve event detection and situational recognition based on transitions in role counts and changes in extracted features.

Circulator with Large Movements

When quantizing features using the role estimation model, the selected features fluctuate for individuals exhibiting significant movement. In emergency situations, assistant personnel move actively, similar to circulators, when preparing for transfusions. We propose Algorithm 3 to count the number of highly active individuals and detect events.

Clustering Latent Feature by K-means

Major changes in surgical situations primarily occur at the start and end of surgery, as well as during emergencies. We analyze activity patterns and tendencies for each role at these key moments. A K-means ($k=5$) model was trained for each role, clustering the latent variables of the proposed model, as shown in Figure 8.7.

8.3.2 Dataset

Table 8.2 presents the dataset used for the End-to-End individual activity recognition model. All datasets consist of surveillance videos covering the entire surgical process,

Algorithm 3 Count the number of circulators with large movements

```

1: The observed data distribution for the entire video is  $X \in \mathbb{R}^D$ , where  $D$  denotes the dimension of data.
2: Extract circulator-classified samples  $X_3 \in X|c = 3$ , where  $c$  denotes the label.
3: Given time  $t$  and the time length of a video  $T$ .
4: Initialize the numbers of circulators with large movements  $N = \{n_0 = 0, \dots, n_T = 0\}$ .
5: Set threshold for the number of different latent values  $z_{th} = 9$ .
6: while  $x \in X_3$  do
7:   Given the start time  $T_s$  and end time  $T_e$  of tracking duration.
8:   while  $T_e < t \leq T_i$  do
9:     Calculate  $z_{q_{t-1}}$  and  $z_{q_t}$  using  $q_\omega(z|x)q_\varphi(c|z)\hat{p}_\theta(z_q|z, c)$ .
10:    if Count of  $i \in \{z_{q_{t-1}, i} \neq z_{q_t, i}\}_i^{15} \geq z_{th}$  then
11:       $n_t = n_t + 1$ 
12:    end if
13:  end while
14: end while
15: Set the event flags  $F = \{f_0 = 0, \dots, f_T = 0\}$ .
16: Set threshold for the number of individuals  $n_{th} = 2$ .
17: while  $n_t \in N$  do
18:   if  $n_t \geq n_{th}$  then  $f_t = 1$ 
19:   end if
20: end while

```

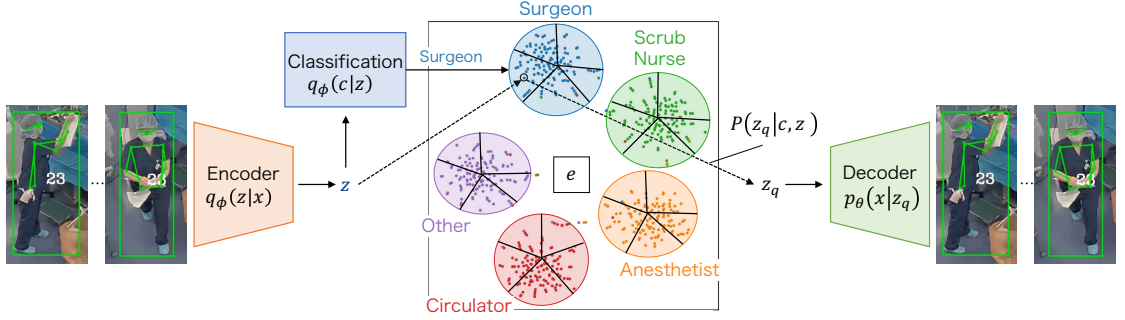


Figure 8.7: Clustering the latent variables of the CSQ-VAE using K-means ($k=5$) for each role. The latent variable z is input into the K-means model for classification.

from preparation to cleanup. These include the scenes used for training the role classification model (Chapter 7). Only Surgery 1 involved an emergency, whereas the others proceeded smoothly until completion.

Table 8.2: Contents of operating room surveillance videos used for event detection and situational recognition through end-to-end individual activity recognition.

	Time Length	Preparing	Proceeding Smoothly	Cleaning	Emergency
Surgery 1	10:00:00	✓	✓	✓	✓
Surgery 2	07:00:00	✓	✓	✓	
Surgery 3	07:30:00	✓	✓	✓	

8.3.3 Prediction

We obtained samples by applying data sampling (Section 7.4.2) and scaling (Section 7.4.3) algorithms, as well as the role classification model trained in Section 7.4.7, to the entire dataset.

8.3.4 Results

Circulator with Large Movements

Figure 8.8 to Figure 8.10 show the results of applying Algorithm 3 to Surgeries 1 to 3. The videos were segmented into 30 min intervals, capturing scenes from the start or end of each surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3. Additionally, Figure 8.11 to Figure 8.13 present frames from the start, end, and key event moments of each surgery.

In Figure 8.8 (Surgery 1), the event flag dropped after the surgery began and remained low until an emergency occurred. Additionally, the event flag rose approx. 5 min after the surgery concluded. As shown in Figure 8.11, more than four individuals were classified as circulators in all frames. Specifically, ID:113 and ID:119 in (a), ID:894, ID:974, ID:1003, and ID:1020 in (b), and ID:381, ID:467, and ID:468 in (c) were moving around the operating room.

In Figure 8.9 (Surgery 2), the event flag dropped after the surgery began and rose again after the surgery ended. The event flag also rose when the surgeon left the surgical field, causing a temporary pause in the procedure. As shown in Figure 8.12 (a), five individuals were classified as circulators. Among them, ID:920 and ID:921 were moving around the operating room, while ID:335 and ID:918 were moving toward the operating table. In Figure 8.12 (b), although ID:185, ID:206, and ID:215 were actually surgeons, they moved away from the operating table and were seen discussing in front of the monitor. During this time, they exhibited movements such as pointing at the monitor. Figure 8.11 (c) shows that at the end of the surgery, surgeons ID:561 and ID:604, who were removing their gowns, were classified as circulators, leading to a rise in the event flag.

In Figure 8.10 (Surgery 3), the event flag dropped 3 min after the surgery began and rose again after the surgery ended. Figure 8.13 (a) and (b) represent the start and end of Surgery 3, respectively. The surgeon and scrub nurse approached the operating table

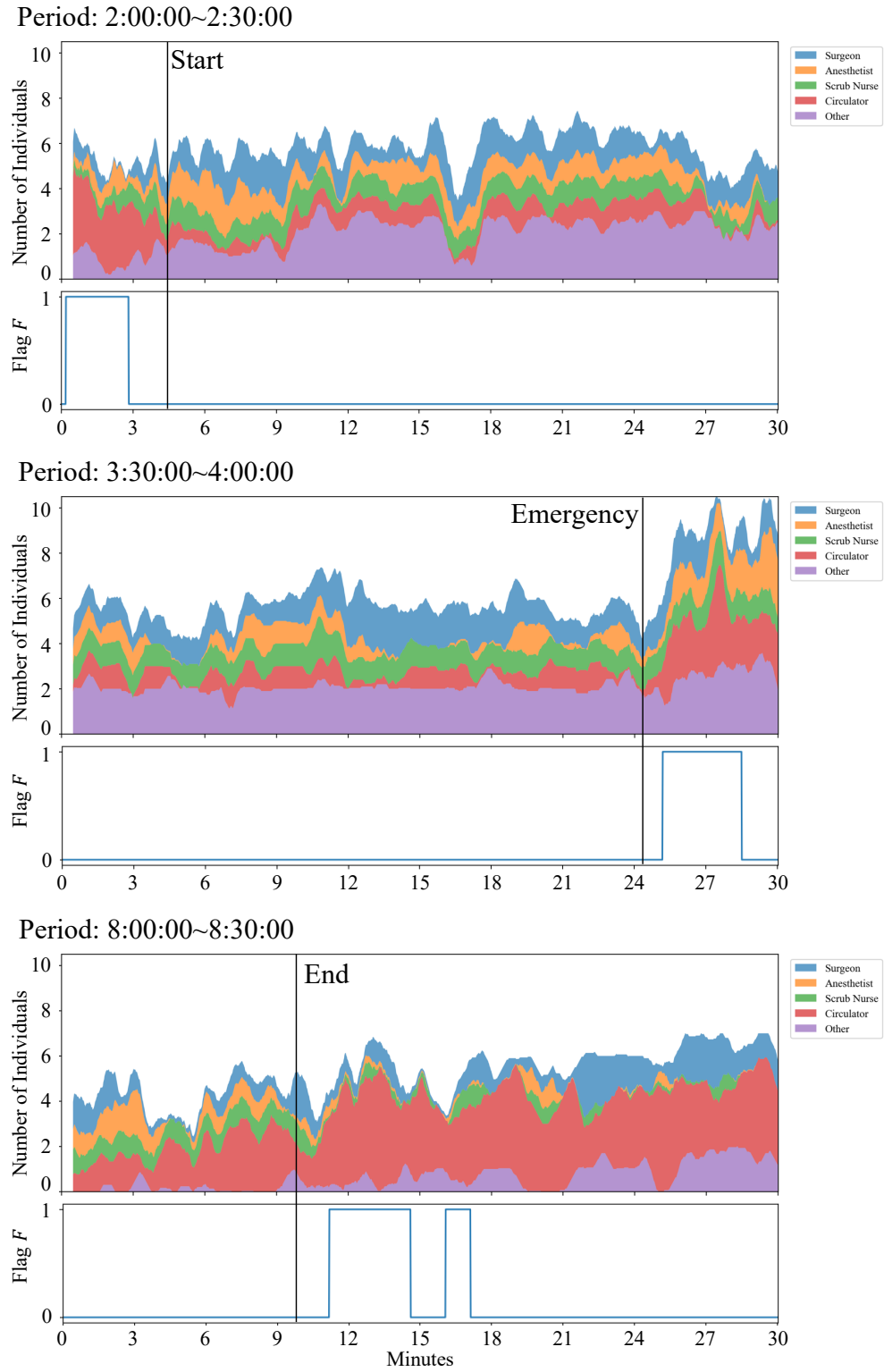


Figure 8.8: Changes in the number of predicted roles and event detection results for Surgery 1. The videos were segmented into 30 minutes intervals, capturing the start and end of surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3.

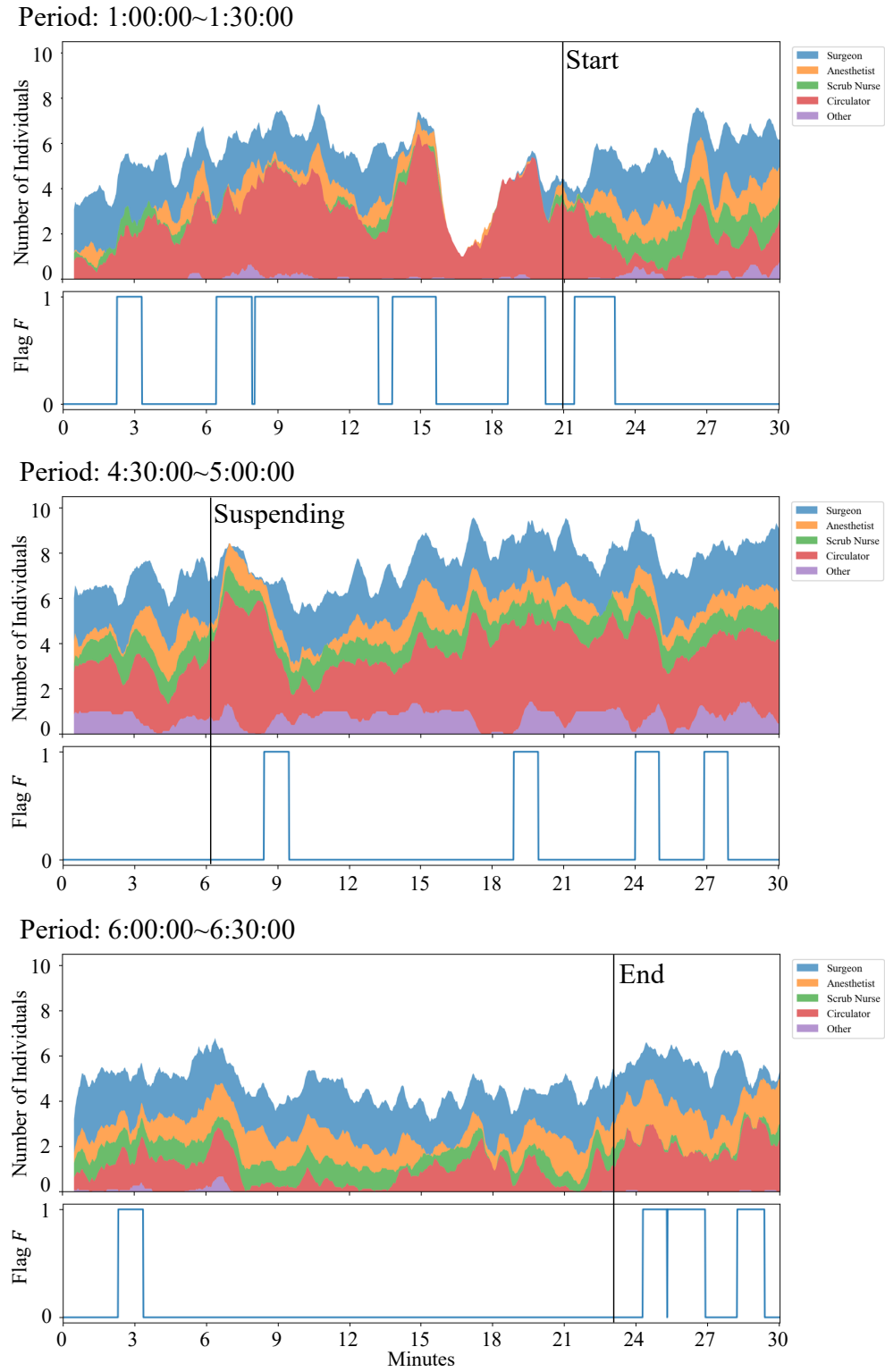


Figure 8.9: Changes in the number of predicted roles and event detection results for Surgery 2. The videos were segmented into 30 minutes intervals, capturing the start and end of surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3.

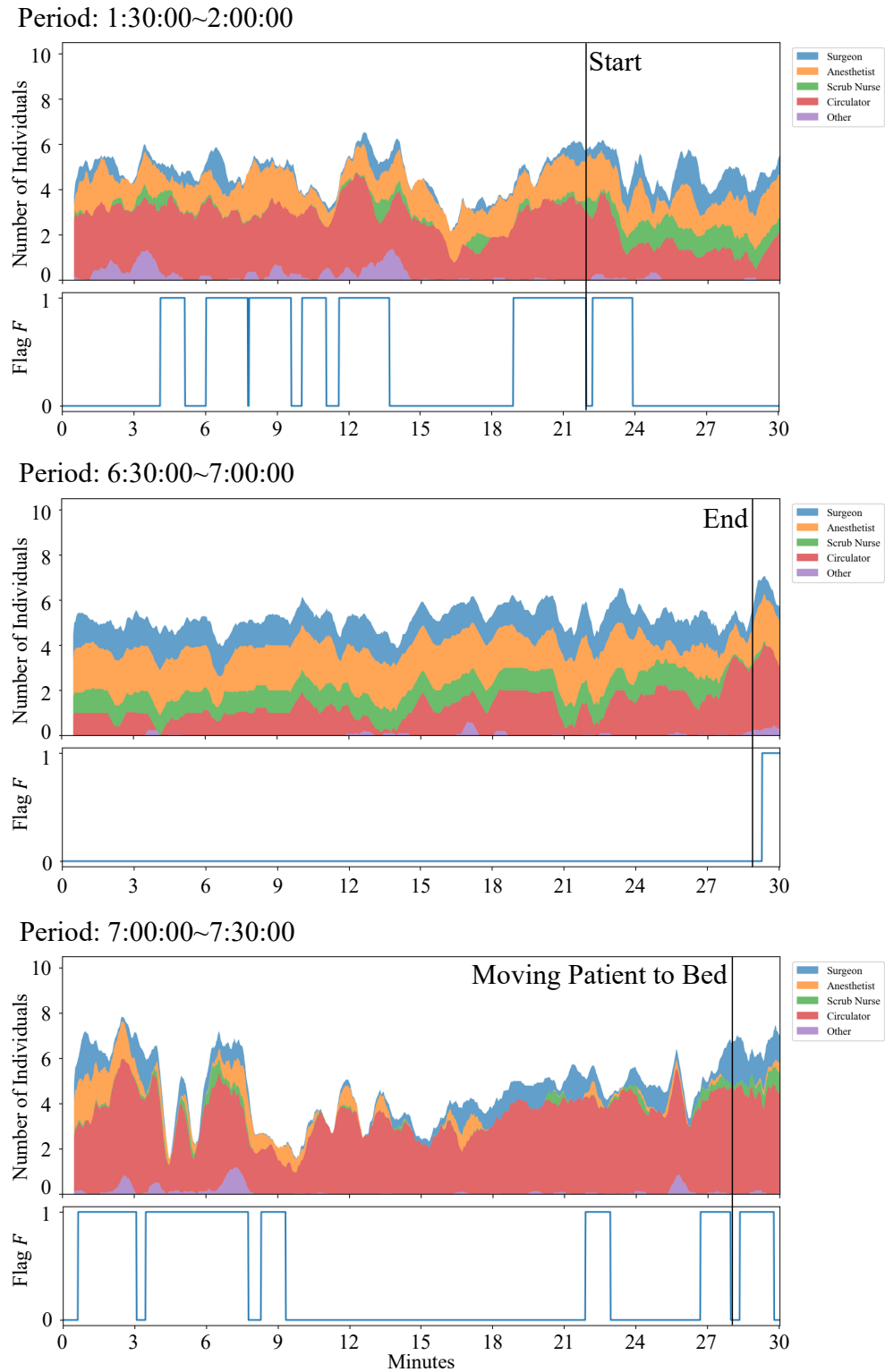


Figure 8.10: Changes in the number of predicted roles and event detection results for Surgery 3. The videos were segmented into 30 minutes intervals, capturing the start and end of surgery as well as other significant events. The “period” represents the time elapsed within the entire surgery. The upper part of each graph illustrates changes in the number of classified roles, while the lower part displays event flags (0: normal, 1: event occurrence) detected by Algorithm 3.



Figure 8.11: Sample frames capturing event occurrence during Surgery 1.

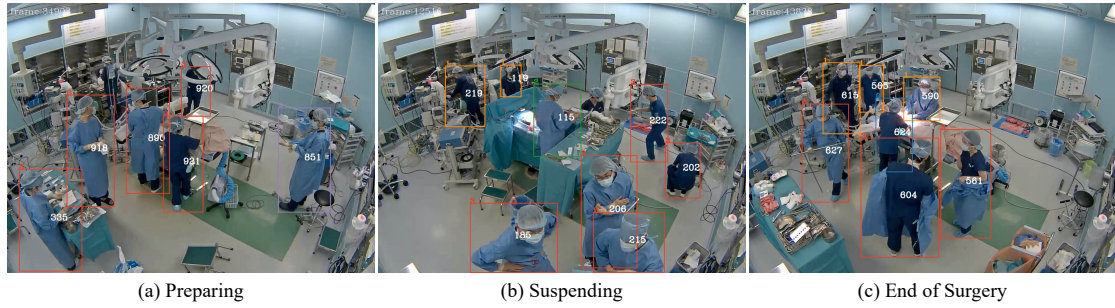


Figure 8.12: Sample frames capturing event occurrence during Surgery 2.

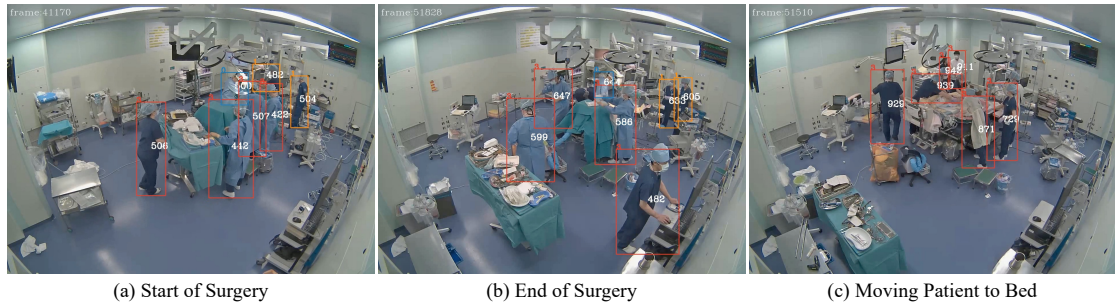


Figure 8.13: Sample frames capturing event occurrence during Surgery 3.

in (a) and moved away from it in (b). In Figure 8.13 (c), the surgical team members made significant movements while transferring the patient from the operating table to the bed.

In situations where the surgery did not proceed smoothly, such as during preparation, cleaning, emergencies, and pauses, the number of individuals classified as circulators who exhibited significant movement in position and posture increased. This suggests the potential for detecting intraoperative changes and events based on the number of circulators and the magnitude of their movements.

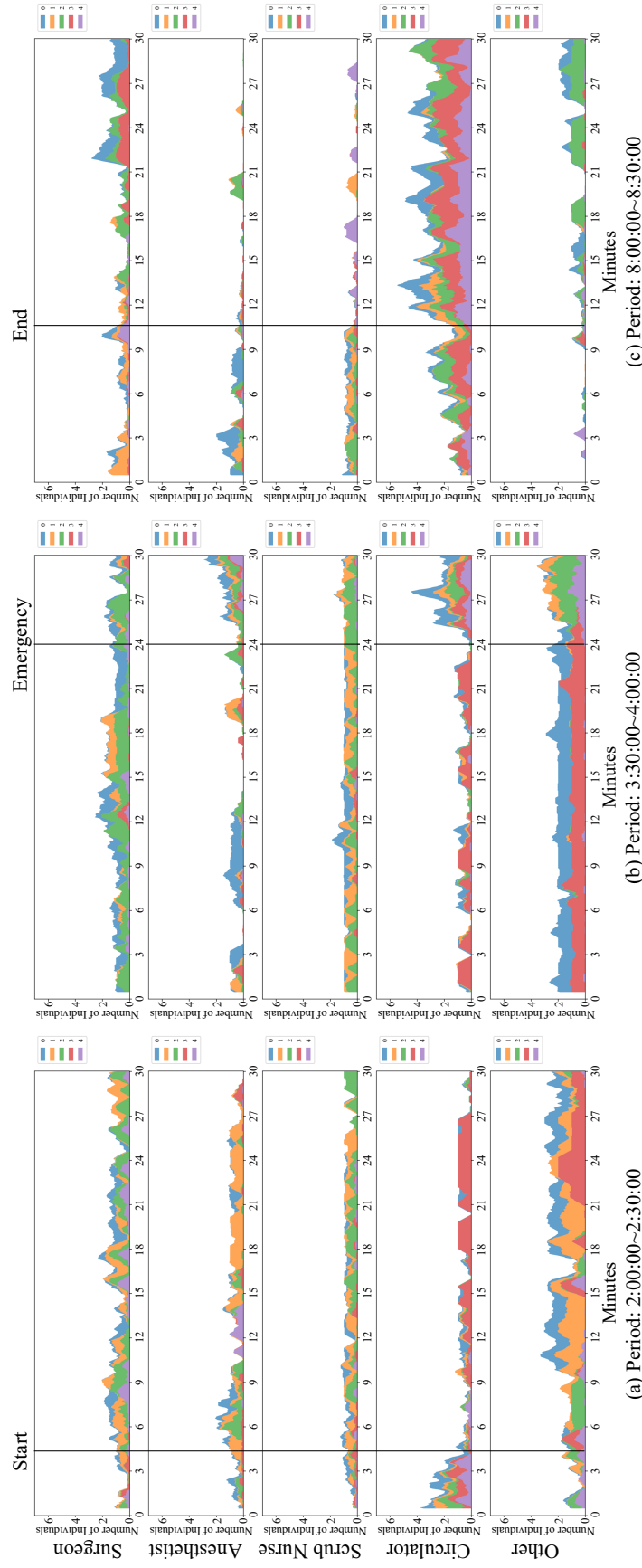


Figure 8.14: Changes in the number of K-means clustering results for Surgery 1. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The "period" represents the time elapsed within the entire surgery.

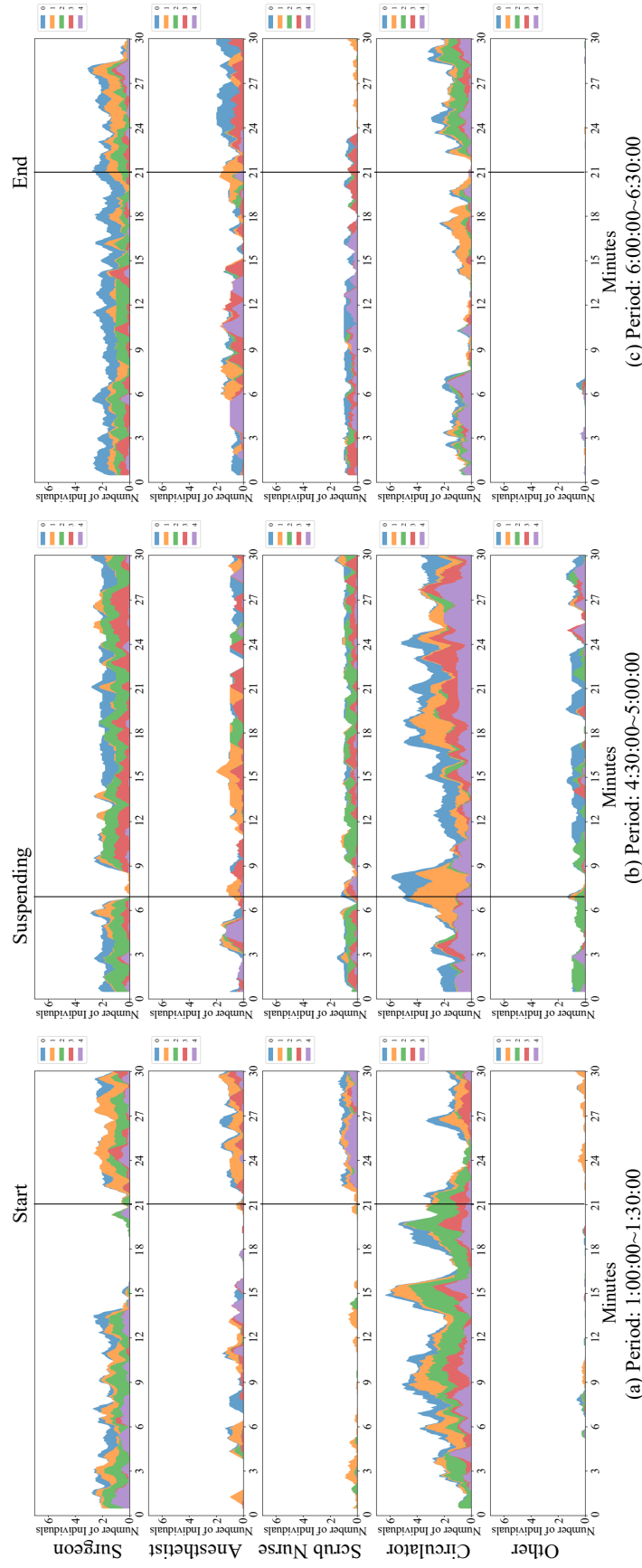


Figure 8.15: Changes in the number of K-means clustering results for Surgery 2. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The "period" represents the time elapsed within the entire surgery.

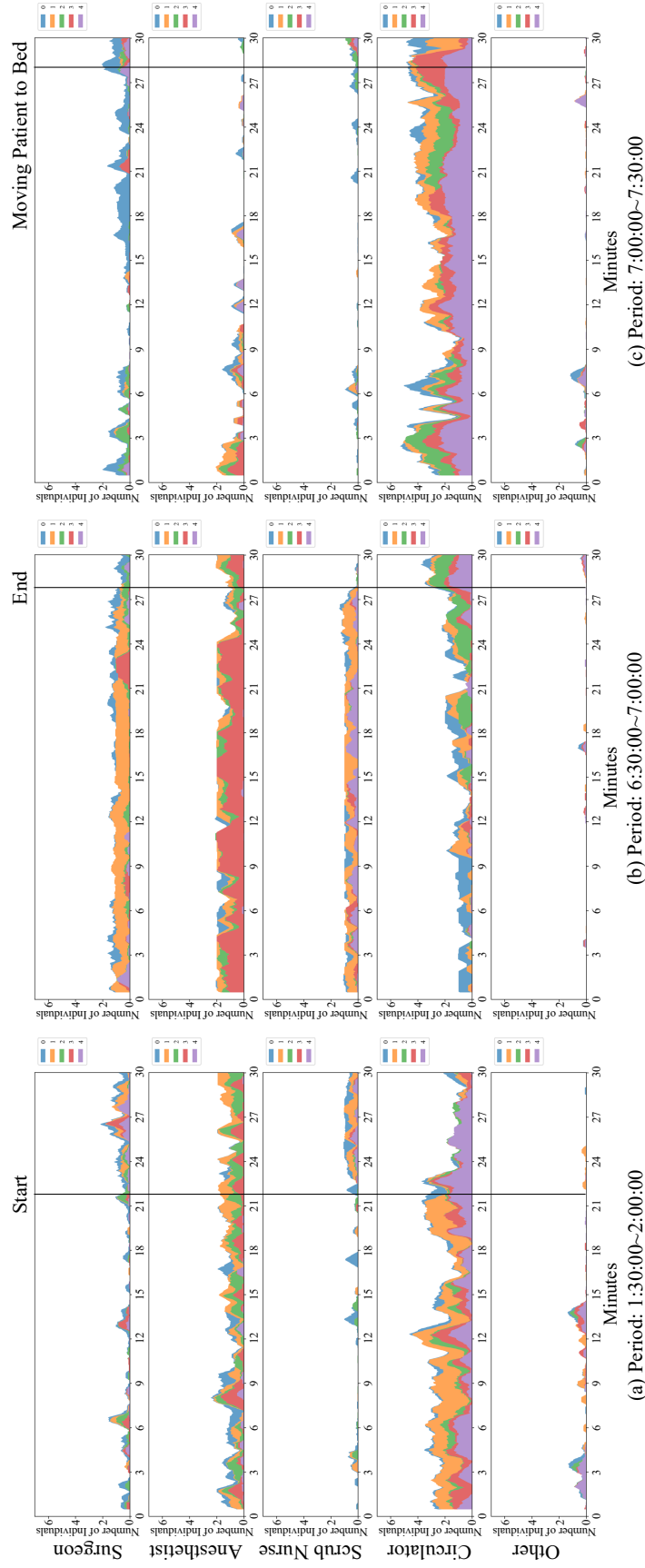


Figure 8.16: Changes in the number of K-means clustering results for Surgery 3. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The "period" represents the time elapsed within the entire surgery.

Clustering Latent Feature by K-means

Figure 8.14 to Figure 8.16 show the results of clustering the model's latent variables using K-means for each role. In each image, the results are arranged from top to bottom in the following order: surgeons, anesthetists, scrub nurses, circulators, and others. The "period" represents the time elapsed within the entire surgery.

Figure 8.14 presents the results of Surgery 1. Circulators were primarily classified into Cluster 3 (red) from the start of surgery until just before the emergency. However, during preparation, the emergency, and cleaning, they were distributed across multiple clusters. This suggests that the activity features of Cluster 3 (red) represent the primary circulator role. Interns, who entered and exited after the surgery began, were clustered into Clusters 0 (blue), 1 (orange), and 3 (red), with a shift in clustering trends before and after the emergency, from Clusters 0 (blue) and 3 (red) to Clusters 1 (orange), 2 (green), and 4 (purple).

Figure 8.15 presents the results for Surgery 2. Among surgeons, the proportion classified into Cluster 1 (orange) was high before and after surgery, while during the mid-surgery period, Clusters 2 (green) and 3 (red) were more prominent. Circulators were more frequently classified into Cluster 2 (green) during preparation and cleaning than during surgery. Additionally, during a suspension event in Figure 8.12 (b), surgeons who moved away from the surgical field and gathered in front of the monitors were classified into Clusters 0 (blue) and 1 (orange).

Figure 8.16 presents the results for Surgery 3. Surgeons were frequently classified into Cluster 1 (orange) during surgery, but after surgery, some were classified into Clusters 0 (blue) and 2 (green). Anesthetists were primarily classified into Cluster 3 (red) during surgery, while their classification into Clusters 1 (orange) and 2 (green) increased in other situations. Circulators were occasionally classified into Cluster 4 (purple) during preparation and cleaning but rarely during surgery.

These findings confirm that the latent variables for each role change according to intraoperative situations and events. Since the role classification model is based on a VAE, the latent variables capture the activity features of each role, which vary depending on surgical events. By clustering these features with K-means, we propose the potential for detailed intraoperative event detection using a neural network model.

8.4 Discussion

In Section 8.2.3, we confirmed that the MIP recognition model successfully recognized activities involving two individuals performing cooperative work, representing direct interactions. Additionally, in Section 8.2.3, the estimated focusing areas of surgical team members helped identify indirect interactions, such as shared attention to the surgical field, which reflects the social structure of the operating room. Social structures can be viewed hierarchically, from individuals as the smallest units to pairs interacting directly and larger groups focusing on a common area. The accumulation of these structures forms the overall social dynamics of the operating room.

In Section 8.3.4, we found that counting the number of circulators making significant movements in position and posture may facilitate situation recognition and event detection. Furthermore, in Section 8.3.4, we confirmed that activity features change according to intraoperative situations and events. By mapping individual activity features to social structures, it may be possible to represent structural changes in the operating room. Representing social structures as graphs, where nodes represent individual activities and edges represent interactions, could enable detailed recognition of situational changes and events.

8.5 Conclusion

In this chapter, we applied the methods proposed in Part I and Part II to operating room surveillance videos covering the entire surgical process, from preparation to cleaning, to recognize intraoperative situations and detect events. From the Part I method, we found that intraoperative situation and event detection could be recognized based on the frequency of activities detected by MIP and changes in the focus areas identified by GA. These factors also reflected interactions between individuals within the operating room. From the Part II role classification model, we found that situation and event detection could be recognized based on the number of significantly moving circulators and changes in the latent variables for each role.

We propose that relationships between individuals, derived from interactions, form the social structure of the operating room. Furthermore, the latent variables extracted by the role classification model represent individual activity features, and mapping them onto the social structure may allow for detecting changes in intraoperative social dynamics.

Chapter 9

Conclusion

9.1 Contributions

Part I Linear Transformation-Based Group Activity Recognition

Chapter 2 Linear Transformation-Based Individual Activity Features

In Chapter 2, we presented a feature extraction method for individual activities using linear transformation. We input time series keypoints obtained from a pose estimation model and a tracking model.

First, we applied homography transformation to keypoints, projecting them onto a top-down view of the operating room. This transformation enables the visualization of feature distributions and changes within the operating room, providing interpretable data. Next, using the keypoints in the top-down view, we applied linear transformation to extract body position, body orientation, and face orientation. Additionally, we extracted arm flexion and wrist position directly from operating room surveillance video (without homography transformation).

Chapter 3 Medical Instrument Passing

In Chapter 3, we proposed a model for recognizing medical instrument passing (MIP) using the individual activity features extracted in Chapter 2. The circulator's activity of passing medical instruments to the scrub nurse serves as an indicator of surgical technique. Since surgical procedures are generally determined by the type of surgery,

monitoring whether instruments are passed at appropriate times allows for the quantification of surgical progress. In the event of an emergency, delays or deviations from the expected instrument handover process can serve as key indicators.

For the experiments, we prepared both simulated and real surgical video datasets to train the model. The model achieved high accuracy in recognizing MIP across both datasets. Furthermore, our proposed method demonstrated potential for recognizing collaborative tasks performed by two individuals facing each other.

Chapter 4 Group Attention

In Chapter 4, we estimated gaze direction based on face orientations extracted in Chapter 2 and identified overlapping areas of attention among surgical team members. During smooth surgical procedures, attention is primarily focused on the operative field. However, during preparation, cleaning, or emergencies, such as when multiple team members enter and exit, attention is more likely to shift away from the operative field. By estimating gaze direction from face orientation and detecting the objects or areas being focused on, we propose that events during surgery can be identified.

For the experiments, we predefined the positions of several objects and quantified the extent to which each object was attended to. The results demonstrated that estimating face orientation effectively identifies the regions or objects of interest to the surgical team. Furthermore, by detecting changes in areas of interest using the proposed method, we found potential for recognizing situational changes or events during surgery.

Part II End-to-End Individual Activity Recognition

Chapter 5 Unsupervised Anomaly Estimation

In Chapter 5, we addressed the estimation of activity anomalies. During smooth surgical procedures, the positions and poses of surgical team members remain largely unchanged. However, in events such as emergencies, active movement increases, such as during preparations for blood transfusions. To estimate anomaly values, we trained a model to reconstruct individuals whose positions and poses are typically stable. Significant reconstruction errors indicate rare activities occurring during surgery. We propose that fluctuations in the proportion of individuals with high reconstruction errors within the operating room can signal event occurrences or situational changes.

We introduced a method for automatically estimating individual activities using anomaly detection based on time series prediction. To evaluate the model's performance, we analyzed two types of time series features: bounding boxes and keypoints. Experimental results showed that while bounding boxes effectively predicted and detected local movements, anomaly scores based on keypoints were largely ineffective for activity detection.

Chapter 6 Unsupervised Activity Clustering

The placement of surgical team members is determined by their roles. For example, the surgeon and scrub nurse are positioned around the operative field and remain largely stationary throughout the surgery. The anesthetist typically stays in front of the anesthesia machine, monitoring the patient, while the circulator moves frequently, transporting medical instruments and handing them to the scrub nurse. Given these patterns, we consider that activity features specific to each role can be extracted and modeled. However, since most surgical team members exhibit limited movement, distinguishing their activities is challenging. As a result, expert annotations are required, leading to high annotation costs.

In Chapter 6, we applied an unsupervised learning model to surgical videos for activity clustering. The results showed that activity clustering achieved over 70% accuracy on the training dataset and over 60% on the test dataset. Moreover, by adjusting the number of clusters (k) and incorporating operating room-specific spatial features, as in Condition 4, accuracy improved by 10%. Under Condition 4, activity classification reached approximately 82.6% accuracy on the training dataset and 70.6% on the test dataset. These findings suggest that spatial movement is a key feature in surgical activity classification, as role placements are fixed.

Chapter 7 Semi-Supervised Role Classification

In Chapter 7, we proposed a model for role classification using semi-supervised learning, integrating the advantages of the models introduced in Chapter 5 and Chapter 6. Bounding boxes and keypoints from pose estimation, represented as coordinate points, preserve spatial movement features while removing background information. This reduces data volume compared to pixel-based representations, making model training and prediction more efficient than in Chapter 6. As a result, the proposed model enables simultaneous activity reconstruction and classification.

Role classification was performed using semi-supervised learning. For surgeries included in the training dataset, the model achieved an accuracy exceeding 0.77 in the test dataset. For surgeries not included in the training dataset, accuracy surpassed 0.85, depending on the arrangement of the operating table within the surveillance camera frames. These results suggest that the proposed model can classify roles by incorporating activity features such as spatial movement magnitude and body orientation. Furthermore, by adding relative features to the input data, such as the position relative to the operating table, the model demonstrated the potential to classify roles even for previously unseen data.

Part III Intraoperative Situational Analysis

Chapter 8 Intraoperative Situation Recognition and Event Detection

In Chapter 8, we applied the methods proposed in Part I and Part II to operating room surveillance videos, covering the entire surgical process from preparation to cleaning, to recognize situations and detect events during surgery.

The application of the method from Part I suggests that changes in the frequency of recognized MIP activities and the areas of attention indicated by GA can be used to identify changes in surgical situations and detect events. Additionally, these activities provide insights into interactions among individuals within the operating room.

The application of the role classification model proposed in Part II indicates that variations in the number of circulators, the intensity of their activities, and changes in latent variables extracted for each role by the model can also serve as indicators of surgical situation changes and event detection.

Summary of Our Achievements

Our research proposed foundational methods of human activity recognition that are intraoperative situation understanding using operating room surveillance video. In Part I, we proposed group activity recognition models for collaborative activities of two individuals and interested areas by surgical team members. In Part II, we proposed individual activity recognition models for role classification and anomaly estimation. In Part III, we applied these activity recognition models to the entire operating room surveillance videos to verify the recognition of surgery start/end and emergencies. Additionally, we confirmed that the features of individual activities for each role change with variations

in the situation. Therefore, our research have potential to be utilized for intraoperative situation understanding.

9.2 Future Direction

9.2.1 Required Research Advancements for Surgical Application

As shown in Section 1.6, we consider that our research has three surgical applications. We will discuss the necessary advancements to realize those applications.

Safety Improvements In Part III, the detection of emergency situations was discussed. To enhance safety in surgical environments, it is essential to accurately detect incidents at the precise timing and recognize the types of incidents occurring.

Surgical Education In Part I and Part II, we proposed role classification models and recognition models for collaborative activities involving two individuals. For effective training, it is necessary to develop models that recognize surgery-specific activities for each role and group activities involving two or more team members. Additionally, to enhance simulations using Virtual Reality (VR) or Augmented Reality (AR), activity generation models are required.

Surgical Evaluation Real-time evaluation metrics are crucial for recognizing surgery-specific activities similarly to surgical evaluation. It is also essential to monitor fine-grained changes in the surgical situation. To achieve this, models that can identify whether activities are appropriate based on the context are necessary. For post-operative evaluation, if the surgery duration was longer than planned, if the blood loss exceeded expectations, or if post-operative complications occurred, methods are needed to extract and analyze potentially causative scenes from surveillance videos based on real-time evaluation metrics.

Therefore, the research developments can be summarized as follows for each category: Human Activity Recognition, Situation Recognition, and Event Detection.

Human Activity Recognition

- **Role-based Activity Recognition:** Recognize activities based on the roles of surgeon, anesthetist, scrub nurse, circulator, and other staff.
- **Group Activity Recognition for Surgical Teams:** Understand the activities of the entire surgical team and relationships among the team members.
- **Activity Generation for Simulation:** Generate activities for simulation purposes.

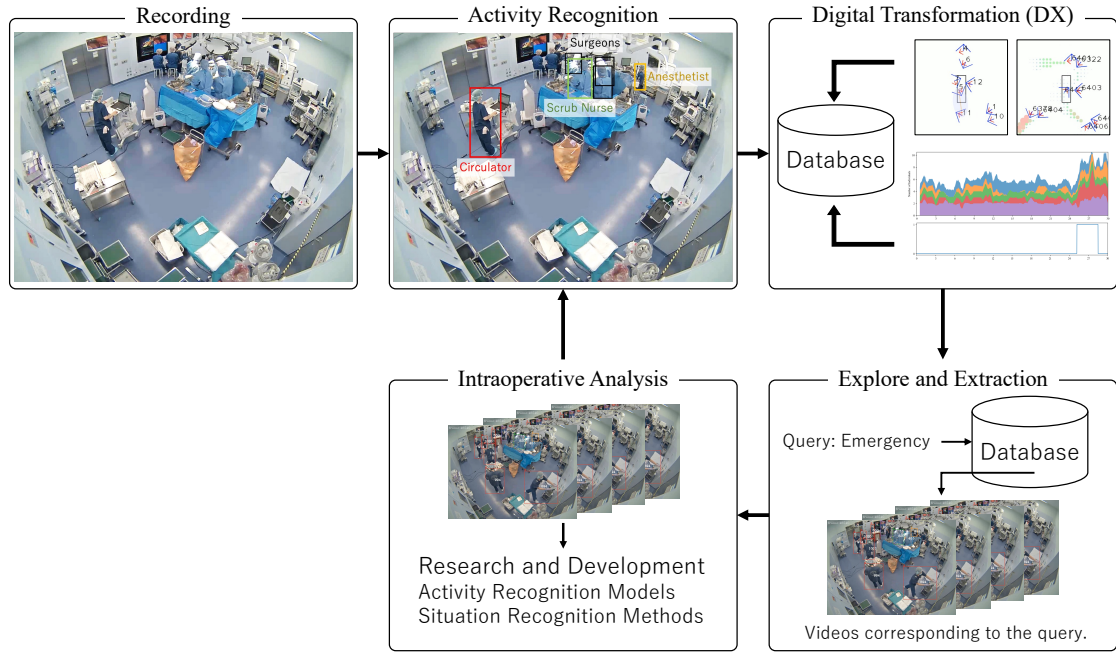


Figure 9.1: Overview of future directions.

Situational Recognition

- Situation Classification: Classify various situations using operating room surveillance video.
- Estimation of Surgical Progress: Estimate the progress of the surgery.

Event Detection

- Precise Event Detection: Detect critical events at the precise timing during surgery.
- Incident Identification: Identify whether the detected event is an incident.

9.2.2 Effective Utilization of Operating Room Surveillance Video

To integrate these approaches into surgical systems for practical applications, further analysis is required, as illustrated in Figure 9.1:

1. Perform activity recognition using the proposed methods.
2. Extract metrics from the activity recognition results and save them, time-stamped with the operating room surveillance video.
3. Retrieve operating room surveillance videos related to critical scenes, such as emergencies, from the database.

4. Develop new activity recognition models and situation awareness methods.

By iterating this cycle, we can identify activities and metrics critical to surgical analysis. Using these key metrics, we aim to develop predictive methods and enhance real-time capabilities for practical applications.

9.2.3 Research Environments

To further advance this research, it is necessary to improve the research environment. By achieving these improvements, the research can be generalized and applied to actual surgical environments.

Automatic Parameter Setting

Implementing a system that automatically sets parameters in response to various conditions and situations in the operating room. To apply the method described in Part I, homography transformation parameters must be defined. In this study, these parameters were set to project the corners of the operating table onto corresponding points in the top view. During surgical preparation, these parameters could be automatically adjusted by detecting the operating table using object recognition and identifying its black surface through image processing.

Handling Occlusion

Establishing methods to overcome obstructions and visual hindrances within the operating room, thereby enabling more accurate data collection. For the method described in Part II, it is necessary to annotate operating room surveillance videos and train the role classification model. Since the model employs semi-supervised learning, annotating all surgical team members is not required. Instead, we propose annotating only those who are clearly visible and not occluded, as this approach can still generate a highly accurate role classification model.

Additionally, this study focused on operating room surveillance videos that exclude large medical instruments and robots to minimize occlusion. However, because the operating table is centrally located, complete occlusion avoidance is challenging. To address this, multiple operating room surveillance cameras should be used to capture an overview from multiple angles.

9.2.4 Multimodel Model

To implement practical applications, multimodal models integrating surgical field cameras, vital sensors, and surgeons' voices should be considered. A surgical field camera provides a detailed view of surgeons' movements, which is particularly useful for estimating surgical progress. Currently, at Kyoto University Hospital, a camera is mounted at the center of the shadowless lamp, but it is frequently obstructed by surgeons' heads, limiting its effectiveness. Similar to surveillance cameras, multiple field cameras should be installed to reduce occlusion.

Vital sensors enable continuous patient monitoring, though they are subject to noise interference. Surgeons' voices can help determine the start and end times of surgeries, but microphones must be installed to capture them effectively. Combining these modalities with operating room surveillance videos would enable more detailed surgical analysis, ultimately contributing to improved surgical safety and efficiency.

References

- [1] Linda Kohn, Janet Corrigan, and Molla Donaldson. *To Err is Human: Building a Safer Health System*. NATIONAL ACADEMIES PRESS, 2000.
- [2] Vincent Charles, Moorthy Krishna, Sarker Sudip K, Chang Avril, and Darzi Ara W. Systems approaches to surgical quality and safety: from concept to measurement. *Annals of surgery*, 239:475–82, 2004. doi: 10.1097/01.sla.0000118753.22830.41.
- [3] Thomas G Weiser, Alex B Haynes, George Molina, Stuart R Lipsitz, Micaela M Esquivel, Tarsicio Uribe-Leitz, Rui Fu, Tej Azad, Tiffany E Chao, William R Berry, and Atul A Gawande. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *The Lancet*, 385, 04 2015. ISSN 0140-6736. doi: 10.1016/S0140-6736(15)60806-6.
- [4] The International Surgical Outcomes Study group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. *BJA: British Journal of Anaesthesia*, 117(5):601–609, 10 2016. ISSN 0007-0912. doi: 10.1093/bja/aew316. URL <https://doi.org/10.1093/bja/aew316>.
- [5] J. F. Calland, S. Guerlain, R. B. Adams, C. G. Tribble, E. Foley, and E. G. Chekan. A systems approach to surgical safety. *Surgical Endoscopy And Other Interventional Techniques*, 16:1005–1014, 2002. doi: 10.1007/s00464-002-8509-3.
- [6] Caprice C. Greenberg, Scott E. Regenbogen, David M. Studdert, Stuart R. Lipsitz, Selwyn O. Rogers, Michael J. Zinner, and Atul A. Gawande. Patterns of communication breakdowns resulting in injury to surgical patients. *Journal of the American College of Surgeons*, 204(4):533–540, 2007. ISSN 1072-7515. doi: <https://doi.org/10.1016/j.jamcollsurg.2007.01.010>.
- [7] Tara J T Kennedy, Glenn Regehr, G Ross Baker, and Lorelei Lingard. Preserving professional credibility: grounded theory study of medical trainees’ requests for clinical support. *BMJ*, 338, 2009. ISSN 0959-8138. doi: 10.1136/bmj.b128.

- [8] Lingard L, Reznick R, Espin S, Regehr G, and DeVito I. Team communications in the operating room; talk patterns, sites of tension, and implications for novices. *Academic Medicine*, 77(3):232–237, 2002.
- [9] Alan H. Rosenstein and Michelle O’Daniel. Impact and implications of disruptive behavior in the perioperative arena. *Journal of the American College of Surgeons*, 203(1):96–105, 2006. ISSN 1072-7515. doi: <https://doi.org/10.1016/j.jamcollsurg.2006.03.027>. URL <https://www.sciencedirect.com/science/article/pii/S1072751506003413>.
- [10] Nick Sevdalis, Melinda Lyons, Andrew N Healey, Shabnam Undre, Ara Darzi, and Charles A Vincent. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Annals of surgery*, 249(6):1047–1051, June 2009.
- [11] A Mishra, K Catchpole, and P McCulloch. The oxford notechs system: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *BMJ Quality & Safety*, 18(2):104–108, 2009. doi: 10.1136/qshc.2007.024760.
- [12] Eleanor R. Robertson, Mohammed Hadi, Lauren J. Morgan, Sharon P. Pickering, Gary Collins, Steve New, Damien Griffin, Peter McCulloch, and Ken C. Catchpole. Oxford notechs ii: A modified theatre team non-technical skills scoring system. *PLOS ONE*, 9:1–8, 03 2014. doi: 10.1371/journal.pone.0090320. URL <https://doi.org/10.1371/journal.pone.0090320>.
- [13] Crossley J, Marriott J, Purdie H, and Beard JD. Prospective observational study to evaluate notss (nontechnical skills for surgeons) for assessing trainees’ non-technical performance in the operating theatre. *The British Journal of Surgery*, 2011.
- [14] Whittaker G, Abboudi H, Khan MS, Dasgupta P, and Ahmed K. Teamwork assessment tools in modern surgical practice: A systematic review. *Surgery Research and Practice*, 2015. doi: 10.1155/2015/494827.
- [15] Guilherme Augusto Silva Surek, Laio Oriel Seman, Stefano Frizzo Stefenon, Viviana Cocco Mariani, and Leandro dos Santos Coelho. Video-based human activity recognition using deep learning approaches. *Sensors*, 23(14), 2023. ISSN 1424-8220. doi: 10.3390/s23146384. URL <https://www.mdpi.com/1424-8220/23/14/6384>.
- [16] Mohanad Babiker, Othman O Khalifa, Kyaw Kyaw Htike, Aisha Hassan, and Muhamed Zaharadeen. Automated daily human activity recognition for video surveillance using neural network. In *2017 IEEE 4th international conference on*

- smart instrumentation, measurement and application (ICSIMA)*, pages 1–5. IEEE, 2017.
- [17] Ahmed Taha, Hala H Zayed, Mohamed E Khalifa, and El-Sayed M El-Horbaty. Human activity recognition for surveillance applications. In *Proceedings of the 7th International Conference on Information Technology*, pages 577–586, 2015.
- [18] Lasitha Piyathilaka and Sarath Kodagoda. Human activity recognition for domestic robots. In *Field and Service Robotics: Results of the 9th International Conference*, pages 395–408. Springer, 2015.
- [19] Danica Hendry, Kevin Chai, Amity Campbell, Luke Hopper, Peter O’Sullivan, and Leon Straker. Development of a human activity recognition system for ballet tasks. *Sports medicine-open*, 6:1–10, 2020.
- [20] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors*, 14(7):11735–11759, 2014.
- [21] Eni Halilaj, Apoorva Rajagopal, Madalina Fiterau, Jennifer L Hicks, Trevor J Hastie, and Scott L Delp. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of biomechanics*, 81:1–11, 2018.
- [22] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [23] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [24] Longteng Kong, Jie Qin, Di Huang, Yunhong Wang, and Luc Van Gool. Hierarchical attention and context modeling for group activity recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1328–1332, 2018. doi: 10.1109/ICASSP.2018.8461770.
- [25] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

- [26] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [27] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [28] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7476–7485, October 2021.
- [29] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6955–6968, 2023. doi: 10.1109/TPAMI.2020.3034233.
- [30] Zhengcen Li, Xinle Chang, Yueran Li, and Jingyong Su. Skeleton-based group activity recognition via spatial-temporal panoramic graph. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 252–269, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73202-7.
- [31] Xue Jiang, Linbo Qing, Jianglan Huang, Li Guo, and Yonghong Peng. Unveiling group activity recognition: Leveraging local-global context-aware graph reasoning for enhanced actor-scene interactions. *Engineering Applications of Artificial Intelligence*, 133:108412, 2024. ISSN 0952-1976. doi: <https://doi.org/10.1016/j.engappai.2024.108412>. URL <https://www.sciencedirect.com/science/article/pii/S0952197624005700>.
- [32] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.
- [33] Kirill Gavriluk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [34] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3261–3269, May 2021. doi: 10.1609/aaai.v35i4.16437. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16437>.

- [35] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13668–13677, October 2021.
- [36] Longteng Kong, Wanting Zhou, Duoxuan Pei, Zhaofeng He, and Di Huang. Group activity representation learning with long-short states predictive transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7267–7281, 2023. doi: 10.1109/TCSVT.2023.3278984.
- [37] Chihiro Nakatani, Hiroaki Kawashima, and Norimichi Ukita. Learning group activity features through person attribute prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18233–18242, June 2024.
- [38] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Articulated clinician detection using 3d pictorial structures on rgb-d data. *Medical image analysis*, 35:215–224, 2017.
- [39] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. A multi-view rgb-d approach for human pose estimation in operating rooms. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 363–372, 2017.
- [40] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Articulated clinician detection using 3d pictorial structures on rgb-d data. *Med. Image Anal.*, 35:215–224, 2017.
- [41] Vinkle Srivastav, Thibaut Issenhuth, Kadkhodamohammadi Abdolrahim, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018.
- [42] Altair da Silva Costa. Assessment of operative times of multiple surgical specialties in a public university hospital. *Einstein (São Paulo)[online]*, 15(2):200–205, 2017. ISSN 2317-6385. URL <https://doi.org/10.1590/S1679-45082017GS3902>.
- [43] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [44] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

- [45] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11212>.
- [46] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nan-ni Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [47] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [48] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [49] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [50] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [51] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8561–8568, Jul. 2019. doi: 10.1609/aaai.v33i01.33018561. URL <https://ojs.aaai.org/index.php/AAAI/article/view/4875>.
- [52] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [53] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL <http://arxiv.org/abs/1705.06950>.

- [54] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
- [55] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [56] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks, 2021.
- [57] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, June 2022.
- [58] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, June 2022.
- [59] Zerui Chen, Yan Huang, Hongyuan Yu, Ke Xue, Binand Han, Yiru Guo, and Liang Wang. Towards part-aware monocular 3d human pose estimation: An architecture search approach. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 715–732, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58580-8.
- [60] Sangbum Choi, Seokeon Choi, and Changick Kim. Mobilehumanpose: Toward real-time 3d human pose estimation in mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2328–2338, June 2021.
- [61] Dong-Hyun Hwang, Suntae Kim, Nicolas Monet, Hideki Koike, and Soonmin Bae. Lightweight 3d human pose estimation network training using teacher-student learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [62] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll,

- and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Trans. Graph.*, 39(4), July 2020. ISSN 0730-0301.
- [63] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [64] Thomas Gatt, Dylan Seychell, and Alexiei Dingli. Detecting human abnormal behaviour through a video generated model. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 264–270. IEEE, 2019.
- [65] Onur Temuroglu, Yasutomo Kawanishi, Daisuke Deguchi, Takatsugu Hirayama, Ichiro Ide, Hiroshi Murase, Mayuu Iwasaki, and Atsushi Tsukada. Occlusion-aware skeleton trajectory representation for abnormal behavior detection. In *Frontiers of Computer Vision: 26th International Workshop, IW-FCV 2020, Ibusuki, Kagoshima, Japan, February 20–22, 2020, Revised Selected Papers 26*, pages 108–121. Springer, 2020.
- [66] Satoshi Suzuki, Yukie Amemiya, and Maiko Sato. Skeleton-based visualization of poor body movements in a child’s gross-motor assessment using convolutional auto-encoder. In *2021 IEEE International Conference on Mechatronics (ICM)*, pages 1–6. IEEE, 2021.
- [67] Zhuocheng Jiang, Ge Song, Yu Qian, and Yi Wang. A deep learning framework for detecting and localizing abnormal pedestrian behaviors at grade crossings. *Neural Computing and Applications*, 34(24):22099–22113, 2022.
- [68] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.
- [69] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. doi: 10.48550/ARXIV.2107.08430. URL <https://arxiv.org/abs/2107.08430>.
- [70] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.
- [71] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

- [72] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, jul 2019.
- [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. doi: 10.48550/ARXIV.1706.03762. URL <https://arxiv.org/abs/1706.03762>.
- [74] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. On position embeddings in BERT. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=onxoVA9FxmW>.
- [75] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [76] Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. Deep embedding network for clustering. In *2014 22nd International Conference on Pattern Recognition*, pages 1532–1537, 2014. doi: 10.1109/ICPR.2014.272.
- [77] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *IJCAI*, pages 1925–1931, 2016.
- [78] Ryan McConville, Raúl Santos-Rodríguez, Robert J Piechocki, and Ian Craddock. N2d: (not too) deep clustering via clustering the local manifold of an autoencoded embedding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5145–5152, 2021. doi: 10.1109/ICPR48806.2021.9413131.
- [79] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <https://proceedings.mlr.press/v48/xieb16.html>.
- [80] Jinghua Wang and Jianmin Jiang. Unsupervised deep clustering via adaptive gmm modeling and optimization. *Neurocomputing*, 433:199–211, 2021. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2020.12.082>. URL <https://www.sciencedirect.com/science/article/pii/S0925231220319949>.

- [81] Jinyu Cai, Shiping Wang, Chaoyang Xu, and Wenzhong Guo. Unsupervised deep clustering via contractive feature representation and focal loss. *Pattern Recognition*, 123:108386, 2022. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2021.108386>. URL <https://www.sciencedirect.com/science/article/pii/S0031320321005665>.
- [82] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: A generative approach to clustering. *CoRR*, abs/1611.05148, 2016. URL <http://arxiv.org/abs/1611.05148>.
- [83] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648, 2016. URL <http://arxiv.org/abs/1611.02648>.
- [84] Xiaopeng Li, Zhourong Chen, Leonard KM Poon, and Nevin L Zhang. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. *arXiv preprint arXiv:1803.05206*, 2018.
- [85] He Ma. Achieving deep clustering through the use of variational autoencoders and similarity-based loss. *Mathematical Biosciences and Engineering*, 19(10):10344–10360, 2022.
- [86] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks, 2016.
- [87] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks, 2019.
- [88] Xiaojiang Yang, Junchi Yan, Yu Cheng, and Yizhe Zhang. Learning deep generative clustering via mutual information maximization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6263–6275, 2023. doi: 10.1109/TNNLS.2021.3135375.
- [89] Xiaotong Zhang, Han Liu, Qimai Li, and Xiao-Ming Wu. Attributed graph clustering via adaptive graph convolution. *CoRR*, abs/1906.01210, 2019. URL <http://arxiv.org/abs/1906.01210>.
- [90] Danyang Zhu, Shudong Chen, Xiuhui Ma, and Rong Du. Adaptive graph convolution using heat kernel for attributed graph clustering. *Applied Sciences*, 10(4), 2020. ISSN 2076-3417. doi: 10.3390/app10041473. URL <https://www.mdpi.com/2076-3417/10/4/1473>.

- [91] Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural deep clustering network. In *Proceedings of The Web Conference 2020*, WWW '20, page 1400–1410, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380214. URL <https://doi.org/10.1145/3366423.3380214>.
- [92] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound, 2016.
- [93] Wongun Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1282–1289, 2009. doi: 10.1109/ICCVW.2009.5457461.
- [94] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image Analysis*, 2749:363–370, 2003.
- [95] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL <https://github.com/ultralytics/ultralytics>.
- [96] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
- [97] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL <https://arxiv.org/abs/1312.6114>.
- [98] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.
- [99] Yuhta Takida, Takashi Shibuya, WeiHsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization, 2022. URL <https://arxiv.org/abs/2205.07547>.
- [100] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

- [101] Edouard Pineau and Marc Lelarge. Infocatvae: representation learning with categorical variational autoencoders. *arXiv preprint arXiv:1806.08240*, 2018.
- [102] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/d523773c6b194f37b938d340d5d02232-Paper.pdf.
- [103] Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, 2017. URL <https://arxiv.org/abs/1702.03118>.
- [104] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [105] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [106] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, 2022.
- [107] Vukasin D Stanojevic and Branimir T Todorovic. Boosttrack: boosting the similarity measure and detection confidence for improved multiple object tracking. *Machine Vision and Applications*, 35(3), 2024. ISSN 0932-8092. doi: 10.1007/s00138-024-01531-5.

List of Publications

Journals

- [J1] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Sho Mitarai, Kazumasa Kishimoto, Yukiko Mori and Tomohiro Kuroda. Semi-Supervised Role Classification from Operating Room Surveillance Videos for Intraoperative Situation Analysis. *International Journal of Activity and Behavior Computing*, 2025. (Accepted)
- [J2] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Sho Mitarai, Kazumasa Kishimoto, Yukiko Mori and Tomohiro Kuroda. Spatio-Temporal Unsupervised Individual Clustering for Operating Room Videos. *Applied Intelligence*. (Under Review)
- [J3] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Kazumasa Kishimoto and Tomohiro Kuroda. Operating Room Surveillance Video Analysis for Group Activity Recognition. *Advanced Biomedical Engineering* 12: 171-181, 2023.
- [J4] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Osamu Sugiyama, Luciano Santos and Tomohiro Kuroda. Recognition of Instrument Passing and Group Attention for Understanding Intraoperative Team States. *Advanced Biomedical Engineering* 11: 37-47, 2022.

International Conferences

- [C1] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Sho Mitarai, Kazumasa Kishimoto, Yukiko Mori and Tomohiro Kuroda. Semi-Supervised Role Classification from Operating Room Surveillance Videos for Intraoperative Situation Analysis. *7th International Conference on Activity and Behavior Computing*, 2025.

- [C2] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Kazumasa Kishimoto, Yukiko Mori and Tomohiro Kuroda. Individual Activity Anomaly Estimation in Operating Rooms Based on Time-Sequential Prediction. *MEDINFO 2023 – The Future Is Accessible*, 284-288, 2024. **Best Student Paper Award (Second Place)**.

Domestic Conferences

- [C3] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Sho Mitarai, Kazumasa Kishimoto, Yukiko Mori and Tomohiro Kuroda. Understanding Surgical Situation Using Operating Room Surveillance Videos and an Individual Activity Recognition Model. *The 44th Joint Conference on Medical Informatics*, 2024.
- [C4] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Sho Mitarai, Kazumasa Kishimoto, Yukiko Mori and Tomohiro Kuroda. Anomaly Estimation and Clustering in Individual Activity from Operating Room Surveillance Videos Using Variational Autoencoder. *The 27th Meeting on Image Recognition and Understanding*, 2024.
- [C5] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Kazumasa Kishimoto and Tomohiro Kuroda. Investigating Utility of Group Activity Recognition Based on Operating Room Surveillance Video Analysis. *The 42th Joint Conference on Medical Informatics*, 2022.
- [C6] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Kazumasa Kishimoto and Tomohiro Kuroda. Operating Room Surveillance Video Analysis for Group Activity Recognition. *Japanese Society for Medical and Biological Engineering*, 2022.
- [C7] Koji Yokoyama, Goshiro Yamamoto, Chang Liu, Osamu Sugiyama, Luciano HO Santos and Tomohiro Kuroda. Recognition of Instrument Passing and Group Attention for Understanding Intraoperative State of Surgical Team. *Japanese Society for Medical and Biological Engineering*, 2021.
- [C8] Koji Yokoyama, Goshiro Yamamoto, Hiroshi Sasaki, Osamu Sugiyama, Luciano Santos, Chang Liu, Shusuke Hiragi, Kazuya Okamoto and Tomohiro Kuroda. Understanding an Operating Room Situation Based on Multi-Person Motion Analysis. *The 65th Annual Conference of the Institute of Systems, Control and Information Engineers*, 2021.