EgoSurgery-Phase: A Dataset of Surgical Phase Recognition from Egocentric Open Surgery Videos

Ryo Fujii[1][0000−0002−9115−8414], Masashi Hatano[1], Hideo Saito[1][0000−0002−2421−9862], and Hiroki Kajita[2]

[1] Keio University, Yokohama, Kanagawa, Japan
{ryo.fujii0112, hatano1210, hs}@keio.jp
[2] Keio University School of Medicine, Shinjuku, Tokyo, Japan
{jmrbx767}@keio.jp

Abstract. Surgical phase recognition has gained significant attention due to its potential to offer solutions to numerous demands of the modern operating room. However, most existing methods concentrate on minimally invasive surgery (MIS), leaving surgical phase recognition for open surgery understudied. This discrepancy is primarily attributed to the scarcity of publicly available open surgery video datasets for surgical phase recognition. To address this issue, we introduce a new egocentric open surgery video dataset for phase recognition, named Egosurgery-Phase. This dataset comprises 15 hours of real open surgery videos spanning 9 distinct surgical phases all captured using an egocentric camera attached to the surgeon's head. In addition to video, the Egosurgery-Phase offers eye gaze. As far as we know, it is the first real open surgery video dataset for surgical phase recognition publicly available. Furthermore, inspired by the notable success of masked autoencoders (MAEs) in video understanding tasks (*e.g.*, action recognition), we propose a gaze-guided masked autoencoder (GGMAE). Considering the regions where surgeons' gaze focuses are often critical for surgical phase recognition (*e.g.*, surgical field), in our GGMAE, the gaze information acts as an empirical semantic richness prior to guiding the masking process, promoting better attention to semantically rich spatial regions. GGMAE significantly improves the previous state-of-the-art recognition method (6.4% in Jaccard) and the masked autoencoder-based method (3.1% in Jaccard) on Egosurgery-Phase. The dataset will be released at `https://github.com/Fujiry0/EgoSurgery`.

Keywords: Surgical video dataset · Surgical phase recognition · Open surgery · Masked autoencoder · Egocentric vision

1 Introduction

Automated analysis of surgical videos is indispensable for various purposes, including providing real-time assistance to surgeons, supporting education, and

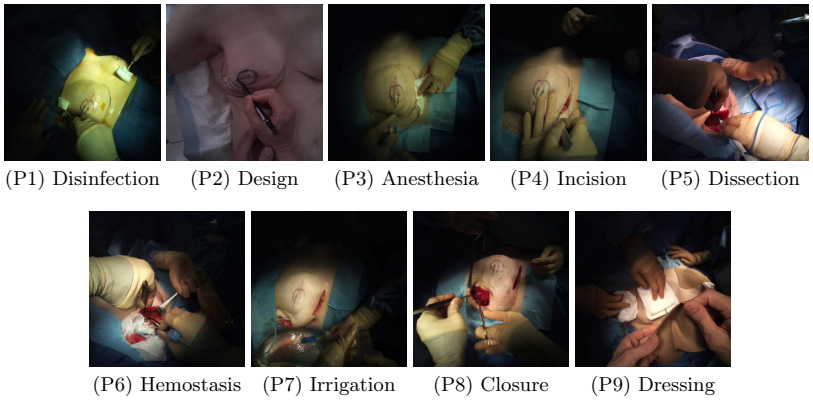

Fig. 1. Illustration of 9 surgical phases (P1-P9) annotated in the EgoSurgery-Phase dataset. Typically, the phases are executed sequentially from P1 to P9.

evaluating medical treatments. Surgical phase recognition, the recognition of the transitions of high-level stages of surgery, is a fundamental component in advancing these objectives. Surgical phase recognition has gained considerable attention with numerous approaches [1, 4, 7, 8, 16, 17, 21]. While surgical phase recognition is important across all surgical methods, the predominant focus of research endeavors has been on minimally invasive surgery (MIS), leaving open surgery phase recognition comparatively underexplored. This discrepancy primarily stems from the scarcity of publicly available large-scale open surgery datasets for phase recognition. In the surgical phase recognition for MIS, several large-scale datasets [17, 20] have been released, driving advancements in learning-based algorithms. Conversely, the absence of comparable large-scale datasets for open surgery phase recognition has significantly impeded progress in achieving accurate surgical phase recognition within the open surgery domain.

To tackle this issue, we introduce Egosurgery-Phase, the first large-scale egocentric open surgery video dataset for phase recognition. 21 videos of procedures of 10 distinct surgical types with a total duration of 15 hours conducted by 8 surgeons are collected and annotated into 9 phases. The videos have been meticulously pre-processed for de-identification. EgoSurgery-Phase offers a rich collection of video content capturing diverse interactions among individuals (*e.g.*, surgeons, assistant surgeons, anesthesiologists, perfusionists, and nurses), varied operative settings, and various lighting conditions. Moreover, in addition to video, EgoSurgery-Phase provides eye gaze data.

Furthermore, inspired by the remarkable performance of Masked Autoencoders (MAEs) [5], which learns meaningful representations by reconstructing the masked tokens, in video understanding tasks (*e.g.*, action recognition), we propose a gaze-guided masked autoencoder (GGMAE). In MAEs, for the selection of masked tokens, a random masking strategy has been often utilized and shown to work well compared to its counterparts in some cases [5, 15, 12]. However, open surgery videos often contained non-informative regions (For in-

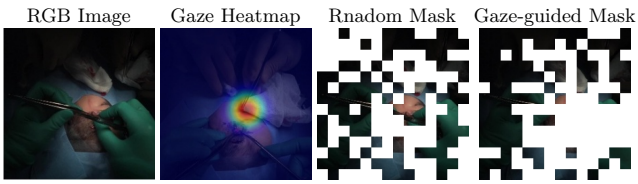

Fig. 2. Example of RGB image and gaze heatmap from EgoSurgery-Phase, along with their corresponding random mask and gaze-guided mask. The gaze heatmap is depicted as a heatmap overlaid onto the RGB image for visualization purposes.

stance, in most sample frames from EgoSurgery-Phase illustrated in Fig. 1, we observe that the intense light from the surgical lamp causes the black clipping to outside the surgical field, making most of the tokens outside surgery field non-informative). Therefore, assuming all tokens have equal information and a uniform probability distribution for masked token selection is suboptimal. With the random masking strategy, masked tokens may be sampled from low-information regions rather than high-information ones, and training to reconstruct these tokens through MAEs is not effective [12, 14]. To address this issue, we propose a gaze-guided masking approach. Given that regions, where surgeons' gaze focuses, are often critical for surgical phase recognition (*e.g.*, the surgical field), our GGMAE leverages gaze information as an empirical semantic richness prior to guiding the masking process, as shown in Fig. 2. It converts input gaze heatmaps into a probability distribution and employs reparameterization techniques for efficient probability-guided masked token sampling. Consequently, tokens that surgeons focus on are masked with higher probability, enabling enhanced attention to semantically rich spatial regions.

Our main contributions are summarized as follows: 1) we constructed the first publicity available large-scale real egocentric open surgery dataset, EgoSurgery-Phase, for phase recognition, 2) we propose a gaze-guided masked autoencoder, GGMAE, which incorporates gaze as an empirical semantic richness prior for masking, and 3) experimental results show that our GGMAE yields significant improvement over existing phase recognition and masked autoencoder-based methods, achieving the state-of-the-art performance on EgoSurgery-Phase.

2 Dataset Design

2.1 Dataset collection

Following the dataset collection protocol proposed in prior research [3], which focused on constructing datasets for surgical tool detection in open surgery videos, we gathered 21 open surgery videos utilizing Tobii cameras attached to the surgeon's head. The recording of patient videos received ethical approval from the Keio University School of Medicine Ethics Committee, and written informed consent was obtained from all patients or their guardians. Our dataset encompasses 10 distinct types of surgeries, performed by 8 different surgeons.

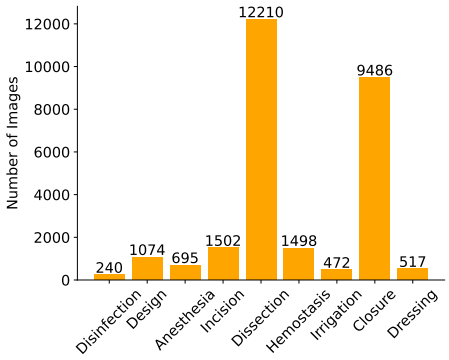

Fig. 3. The phase distribution of frames.

The 21 videos were recorded at a frame rate of 25 fps and a resolution of 1920×1080 pixels. Video durations vary between 28 and 234 minutes, reflecting the diversity in type and complexities of surgery. In total, 28 hours of surgical footage were captured. Unlike videos of minimally invasive surgery (MIS), open surgery videos are more likely to contain personally identifiable information (PII) such as the faces of patients, assistant surgeons, and nurses. To address privacy concerns, we subsampled the videos to 0.5 fps and anonymized the patient's face through blurring. In addition, we exclude frames containing other PII. After these pre-processing steps, the average duration of the videos becomes 46 minutes, resulting in a total duration of 15 hours, thereby yielding a large-scale dataset of high quality. In addition to video, EgoSurgery-Phase provides eye gaze.

2.2 Dataset annotation, statistics and data split

Expert surgeons perform the annotations based on their clinical experience and domain knowledge. The 21 pre-processed videos of open surgery are manually annotated into 9 phases: Disinfection, Design, Anesthesia, Incision, Dissection, Hemostasis, Irrigation, Closure, and Dressing. Samples are shown in Fig. 1. In total, $27,694$ frames are manually annotated. The sample distribution is shown in Fig.3. It reveals a notable class imbalance. We use 14 videos for the training set, 2 videos for the validation set, and 5 videos for the test set.

3 Approach

3.1 Overview

Fig. 4 presents an overview of the proposed GGMAE. GGMAE takes as input video $V \in \mathbb{R}^{T \times C \times H \times W}$ and gaze heatmaps $G \in \mathbb{R}^{T \times H \times W}$. Here, C represents the input (RGB) channels, and $H \times W$ denotes the spatial resolution of each

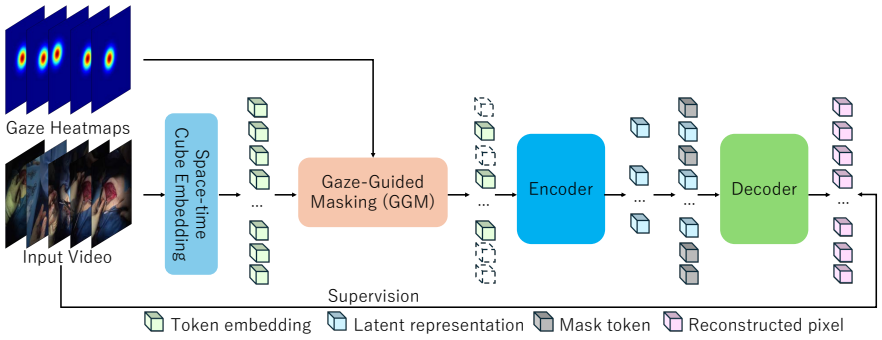

Fig. 4. Overview of the proposed GGMAE: GGME performs the task of masking tokens and reconstructing these masked tokens with Transformer encoder-decoder architecture. Considering that open surgery videos often contain non-informative regions, we introduce the Gaze-Guided Masking (GGM) module, which selects tokens to be masked based on gaze information.

frame. The space-time cube embedding [15] is used to transform the input video into a set of token embeddings $X \in \mathbb{R}^{F \times N_s N_r}$, where F is the channel dimension of the tokens, and $N_s = HW/H_cW_c$ and $N_r = T/T_c$ are the numbers of tokens along the spatial and temporal dimensions, respectively. T_c, H_c, and W_c represent the size of each token along the temporal, height, and width dimensions, respectively.

We apply the proposed Gaze-Guided Masking (GGM) strategy to select tokens for masking with a masking ratio ρ, leveraging the gaze information. The remaining tokens, along with the space-time position embeddings, are fed into the Transformer encoder and decoder [18] to reconstruct the masked maps.

3.2 Gaze-guided mask Masking

Open surgery videos often contain non-informative regions, and training a model to reconstruct these tokens using MAE does not improve model performance [12, 14]. Therefore, inspired by representation learning approaches that leverage MAEs with non-uniform masking tailored to token informativeness across diverse domain data inputs [9, 10, 12–14], we integrate gaze information as an empirical semantic richness prior to guide the masking of embedding features. Specifically, we propose non-uniform token sampling based on the accumulated gaze heatmap value of each token.

First, we compute the accumulated gaze heatmap value d_i for each token $\boldsymbol{x}i \in X$ by summing the heatmap values across the pixels belonging to the token as follows:

$$d_i = \sum_{j \in \Omega_i} G_i, \tag{1}$$

where Ω_i denotes the set of pixels in the gaze heatmap corresponding to the i-th token. We then calculate the masking probability vector $\boldsymbol{\pi}^t \in \mathbb{R}^{N_s}$ for each

token's time index using the softmax function as follows:

$$\boldsymbol{\pi}^t = \mathrm{Softmax}(\boldsymbol{d}^t/\tau), \tag{2}$$

where $\boldsymbol{d}^t \in \mathbb{R}^{N_s}$ represents a vector of accumulated gaze heatmap for each time index t, and τ is a hyper-parameter controlling the sharpness of the softmax function. Finally, the indices of the masked tokens are determined by sampling from a Multinomial distribution with probabilities $\boldsymbol{\pi}^t$, for $\lfloor \rho N_s \rfloor$ trials without replacement for each time index t.

3.3 Loss function

The loss function is the mean squared error (MSE) loss between the input pixel values and the reconstructed pixel values:

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{p \in \Omega} |I(p) - \hat{I}(p)|^2, \tag{3}$$

where p is the masked token index, Ω is the set of masked tokens, I represents the input ground truth frames, and $\hat{I}$ stands for the reconstructed frames.

4 Experiments

4.1 Implementation Details

Network Architecture. We employ the VideoMAE with the ViT-Small [2] backbone. Following VidoeMAE [15], we use the same input patch size of $2 \times 16 \times 16$ ($T_c \times H_c \times W_c$) for all models. We utilize 10-frame clips (T) as input, maintaining a fixed spatial resolution of 224×224 ($H \times W$) across all experiments. To generate the ground-truth gaze heatmaps, we place a Gaussian centered on the ground truth gaze point.

Pre-training details. During pre-training, the masking ratio of the input token is set to 90%. We adopt the AdamW [11] optimizer with a weight decay of $1e^{-4}$ and betas of (0.9, 0.95). We pre-train the network for 800 epochs with a batch size of 256. The learning rate is linearly increased to $1e^{-3}$ from 0 in the first 20 warmup epochs and then decreased to $1e^{-4}$ by the cosine decay schedule. We set the temperature hyperparameter τ to 0.5. The experiments are conducted using the PyTorch framework on three NVIDIA TITAN RTX GPUs.

Fine-tuning details. After the pre-training, we perform fine-tuning. An MLP head is attached to the pre-trained backbone and the whole network is fully fine-tuned for 100 epochs with cross-entropy loss and a batch size of 64. The learning rate is linearly increased to $5e-4$ from 0 in the first 5 warm-up epochs and then decreased to $5e-5$ by the cosine decay schedule. To mitigate class imbalance during fine-tuning, we employ a resampling strategy. All hyperparameters are determined through standard coarse-to-fine grid search or step-by-step tuning.

Table 1. Performance comparison with baseline and state-of-the-art phase recognition models on EgoSurgery-Phase.

Methods	Backbone	Precision	Recall	Jaccard
PhaseLSTM [16]	AlexNet	36.3	33.1	21.9
PhaseNet [17]	AlexNet	37.0	25.7	19.7
TeCNO [1]	ResNet-50	47.7	39.2	27.3
Trans-SVNet [4]	ResNet-50	41.8	35.9	23.1
NETE [21]	Inception v3	43.7	35.2	27.5
GGMAE (Ours)	ViT-S	**51.7**	**45.6**	**33.9**

Table 2. Performance comparison with state-of-the-art masked autoencoder-based models on Egosurgery-Phase. The supervised baseline is ViT-S trained from scratch on Egosurgery-Phase.

Methods	Backbone	Masking	Precision	Recall	Jaccard
Supervised	ViT-S		47.9	31.6	27.1
VideoMAE [15]	ViT-S	Tube masking	49.3	41.6	29.8
VideoMAE V2 [19]	ViT-S	Dual masking	**54.2**	43.2	30.8
SurgMAE [6]	ViT-S	Spatio-temporal masking	52.2	41.9	27.8
GGMAE (Ours)	ViT-S	Gaze-guided masking	51.7	**45.6**	**33.9**

4.2 Evaluation metrics

To quantitatively analyze the performance of our method, we use three widely used benchmark metrics for surgical phase recognition: precision, recall, and Jaccard index. Due to phase class imbalance inherent within the EgoSurgery-Phase dataset, the performance will be reported in macro-average. Macro-average is used in imbalanced multi-class settings as it provides equal emphasis on minority classes.

4.3 Phase recognition performance comparison

Comparison with phase recognition methods: We first compare our approach with current state-of-the-art phase recognition methods, including TeCNO [1], Trans-SVNet [4], and NETE [21], alongside common baselines PhaseLSTM [16] and PhaseNet [17]. The performance of all methods is summarized in Table 1. Our GGMAE notably surpasses the baselines in all metrics. Specifically, our method exhibits a substantial improvement over NETE, which is the best performance among previous state-of-the-art methods, by 8.0% (from 43.7% to 51.7%) in the Precision, 10.4% (from 35.2% to 45.6%) in the Recall, and 6.4% (from 27.5% to 33.9%) in the Jaccard index.

Table 3. Ablation studies on Egosurgery-Phase. We use ViT-S as a backbone for all the experiments.

(a) Mask sampling strategy.

Strategy	Ratio	Jaccard
Random [5]	0.75	28.9
Random [5]	0.90	30.6
Tube [15]	0.90	29.8
Gaze-guided	0.90	**33.9**

(b) Masking ratio (ρ)

Ratio	Jaccard
0.95	31.2
0.90	**33.9**
0.85	31.6
0.80	31.5

(c) Temperature parameter (τ).

τ	Jaccard
1.00	30.1
0.75	30.6
0.50	**33.9**
0.25	27.2

Comparison with masked autoencoder-based methods. After being pretrained with the proposed GGMAE framework, the model exhibits significant performance improvements compared to the model trained from scratch (6% improvement in the Jaccard index). We then compare current state-of-the-art MAE-based methods, namely VideoMAE and VideoMAEV2. Additionally, we evaluate our approach against SurgMAE, which first demonstrates the effectiveness of MAEs in the surgical domain. The performance of all methods is summarized in Table 2. Employing the same backbone and training schema, GGMAE surpasses VideoMAE by 4.1% and VideoMAEV2 by 3.1% and SurgMAE by 6.1% in terms of Jaccard index.

4.4 Ablation study

Mask sampling strategy. To verify the effectiveness of the proposed gaze-guided masking strategy, we compare its performance with that of random and tube masking. As we can see, our gaze-guided masking strategy brings absolute performance improvements of 3.3%. This suggests that the gaze information, as an empirical semantic richness prior, can effectively guide the masking process.
Masking Ratio. As shown in Tab 3 (b), we experimented with different masking ratios. Results show that either too large or too small masking ratios have a negative impact on performance. We empirically found that a masking ratio of 90% exhibits the best results.
Temerature parameter. We experimented with different temperature parameters τ. As the temperature parameter decreases, the region toward which the gaze is directed becomes more likely to be masked. As shown in Tab 3 (c), Our GGMAE exhibits the best performance when temperature parameters τ is 0.5. Overall, a temperature parameter τ is set to 0.5 by default.

5 Conclusion and Future Work

In this paper, we construct the first egocentric open surgery video dataset, Egosurgery-Phase, for phase recognition. We also propose a gaze-guided masked autoencoder, GGMAE, to promote better attention to semantically rich spatial

regions using gaze information. Furthermore, GGMAE achieves substantial improvements compared to the existing phase recognition methods and masked autoencoder methods. The remaining challenges for this dataset involve improving model performance on the Egosurgery-Phase. By releasing this dataset to the public, we, alongside the wider research community, aspire to address these challenges in the future collaboratively. Moreover, we intend to enrich this dataset by augmenting the video content and incorporating footage captured from various perspectives (*e.g.*, assistant surgeons, anesthesiologists, perfusionists, and nurses) to advance the automated analysis of open surgery videos.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 22H03617. We would like to thank the reviewers for their valuable comments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: MICCAI (2020)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR (2021)
3. Fujii, R., Hachiuma, R., Kajita, H., Saito, H.: Surgical Tool Detection in Open Surgery Videos. Applied Sciences (2022)
4. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. In: MICCAI (2021)
5. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: CVPR (2022)
6. Jamal, M.A., Mohareri, O.: SurgMAE: Masked Autoencoders for Long Surgical Video Analysis (2023)
7. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. TMI (2018)
8. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. TMI (2021)
9. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. In: NeurIPS (2022)
10. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., Wang, J.: MST: Masked Self-Supervised Transformer for Visual Representation. In: NeurIPS (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)

12. Mao, Y., Deng, J., Zhou, W., Fang, Y., Ouyang, W., Li, H.: Masked motion predictors are strong 3d action representation learners. In: ICCV (2023)
13. Min, C., Xiao, L., Zhao, D., Nie, Y., Dai, B.: Occupancy-MAE: Self-Supervised Pre-Training Large-Scale LiDAR Point Clouds With Masked Occupancy Autoencoders. IV (2023)
14. Sun, X., Chen, P., Chen, L., Li, C., Li, T.H., Tan, M., Gan, C.: Masked Motion Encoding for Self-Supervised Video Representation Learning. In: CVPR (2023)
15. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS (2022)
16. Twinanda, A.P., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Single- and Multi-Task Architectures for Surgical Workflow Challenge at M2CAI 2016 (2016)
17. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE TMI (2017)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: NeurIPS (2017)
19. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In: CVPR (2023)
20. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: AutoLaparo: A New Dataset of Integrated Multi-tasks for Image-guided Surgical Automation in Laparoscopic Hysterectomy. In: MICCAI (2022)
21. Yi, F., Yang, Y., Jiang, T.: Not End-to-End: Explore Multi-Stage Architecture for Online Surgical Phase Recognition. In: ACCV (2023)