2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385579

SPHASE: Multi-Modal and Multi-Branch Surgical Phase Segmentation Framework based on Temporal Convolutional Network

Jun Long[‡], Junkun Hong[†], Zidong Wang[†], Tingxuan Chen[†], Yunfei Chen[‡], Liu Yang[*†]
[†]School of Computer Science and Engineering, Central South University, Changsha, China
[‡]Big Data Institution, Central South University, Changsha, China

***Abstract*—Surgical phase segmentation plays an important role in computer-assisted surgery systems, aiming to recognize what step or what action is operating in the video frame. Existing methods focus on improving the accuracy and precision of video segmentation, but ignore semantic consistency and temporal continuity of video frames in the intra-phase, which is necessary to apply in realistic computer-assisted equipment. Meanwhile, recent works almost extract long-term dependencies by Temporal Convolutional Network, but we heed high layers in TCN lose fine-grained information for detecting surgical steps and further affect phase segmentation task. To address these problems, we propose a Surgical Phase Segmentation Framework (SPHASE) which contains a multimodal feature fusion process and follows a multi-branch predictor. Moreover, we design a multimodal feature fusion mechanism when aggregate optical flow feature and I3D feature. The extensive experiments on AutoLaparo, Cholec80, and M2CAI2016 datasets demonstrate our method outperforms the state-of-the-art method by a large margin, especially in the JACC metric, which means SPHASE is more applicable in the surgical operating room.**

***Index Terms*—surgical phase segmentation, multimodal feature fusion, temporal convolutional network, attention mechanism**

I. Introduction

The automatic computer-assisted surgical system has proved an outstanding contribution to support technical solutions for an operating room [14]. Especially, by monitoring the surgical processes in videos, the system generates early warning of potential surgical motion deviations and mistakes [39]. Meanwhile, it can help generate surgical decisions, clinical reports, and data annotation [38] which is closely related to patient safety. Furthermore, another significant influence is advancing surgeon skill evaluation [32] and surgeon educational training. However, the surgical phase segmentation task is more complicated because surgical scenes have less inter-phase variance but more intra-phase variance, which is completely contrary to other videos [18]. For example, existing segmentation algorithms for YouTube videos most focus on the transfer of background scenes but surgical video scenes are always unchanged. In addition, the accuracy of frame detection affects video segmentation, while surgery perception is always blurred and occluded by smoke and blood during the operating process [2].

Some research [20] [23] notices that temporal convolutional network (TCN) can capture long-term and large-scale temporal features for surgical phase segmentation task. Yueming Jin et al. [16] propose an end-to-end temporal memory relation network to extract long-term dependency. Sanat Ramesh et al. [26] propose a multi-task multi-stage temporal convolutional network to predict the phase and step during the laparoscopic gastric bypass procedures. Bokai Zhang et al. [36] propose a multimodal-based method including I3D and optical flow to utilize spatial information and temporal information for surgical workflow recognition. Fangqiu Yi et al. [34] propose a not end-to-end training strategy and get top performance on accuracy and precision. Unfortunately, all of them seldom improve the performance on the Jaccard index (JACC), which is calculated by the consistency of intra-phase and is important for computer-assisted surgical systems. For example, automatic surgical robots must receive continuous operating commands through surgical video. In our observation, the performance of JACC is hindered because high layers in TCNs are inevitable coarse access to fine-gained information inevitably. To solve the problem, our solution is both extracting long-term dependencies and enhancing the short-term features in the meantime to keep frame-wise consistency.

In this paper, we propose a novel Surgical Phase Segmentation Framework (SPHASE), that contains three components: the Multimodal Feature Fusion process (MFF), the Surgical Phase Recognition branch (SPR), and the Phase Boundary Regression branch (PBR). In the MFF process, we collect optical flow data from adjacent video frames and extract optical flow features. At the same time, we extract the I3D features from frame-wise images directly. To ensure better fusion between the two features, inspired by contrast learning [35], we design a multimodal feature fusion mechanism to aggregate I3D features and optical flow features. In the SPR branch, we present a variant of Attention TCN [21] [29] to extract the long-term dependencies, which cross-stack the attention layer behind each temporal layer. In the PBR branch, we only append the attention layer to the first and last TCN layers, aiming to enhance short-term feature representation. According to this multi-modality and multi-branch framework, we not only keep the accuracy and precision of surgical segmentation task but also improve the JACC result by a large margin. The quantitative results also prove our framework maintains consistency in the intra-phase surgical phase. In summary, our main contributions are as follows:

[*] Liu Yang is the corresponding author.

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

- **Task contribution.** We thoroughly investigate the surgical video segmentation task and figure out a motivation that JACC results depend on fine-grained information together with long-term and short-term dependencies. The proposed task is a key point in a computer-assisted surgical system which ignored by recent work.
- **Technical contribution.** We design a novel Surgical Phase Segmentation Framework for surgical video segmentation. It aggregates the I3D feature and optical flow feature, following the Surgical Phase Recognition branch and the Phase Boundary Regression branch. Through two combination approaches, we design two different variants of attention and temporal neural network. One is stacking attention layers followed by temporal layers for getting long-term information. Another is introducing attention layers in the first and last temporal layers for reserving short-term information. We conduct extensive experiments on three challenging datasets: AutoLapora, Cholec80, and M2CAI16. The results prove our SPHASE framework outperforms state-of-the-art methods in the JACC metric by a large margin.
- **Community contribution.** Our codes will be ready to open-source on GitHub after the paper is accepted. We hope our work may become one of the enablers for the valuable but relatively unexplored topic and the learning to surgical video analysis community.

II. Related Work

In this section, we elaborate on the existing methods of surgical phase segmentation. We discuss them into four categories: manual feature networks, single-stage methods, multi-branch framework, and multimodal-based approach.

A. Manual feature networks

The early research on surgical phase segmentation utilizes handcrafted features to represent the information of each video frame, such as value, shape, texture, color, etc [7]. The low-level features cannot express the complicated surgical video since the surgical procedure has little change during the two phases. Therefore, the performance of manual feature networks is limited because pre-defined labels are not sufficient to accurately represent the surgical operation with strongly nonlinear.

B. Single-stage methods

This line of research recognizes the surgical phase based on a video feature extractor and a deep neural network. Dynamic time warping, conditional random field, and Hidden Markov models are commonly used in video feature extractors.

As one of the earliest representative studies, Andru Twinanda et al. [30] first propose a deep learning method for surgical workflow analysis including phase segmentation and tool detection. Thereafter, Sebastian Starke et al. [27] design a baseline method with a convolutional neural network and a recurrent neural network. However, the performance of surgical video analysis is still quite poor. Andru Putra Twinanda [31] extracts long-term dependencies from the surgical video. Toward this goal, they introduce a Long-Short Term Memory (LSTM) for temporal refinement. Similarly, Robert DiPietro et al. [9] train an end-to-end RNN model to learn the temporal correlation of consecutive frames, which uses ResNet to extract features and applies an LSTM network to learn the temporal information. Unfortunately, the average duration of surgical video grows into many hours, which is challenging for LSTM-based methods to leverage the temporal information. To achieve this objective, they introduce a Long-Short Term Memory (LSTM) for temporal refinement. EndoLSTM extracts long-term dependencies from the surgical video.

Beyond that, Temporal Convolutional Networks [10] are proposed for action segmentation in hierarchical videos. In contrast to RNN, the encoder-decoder structure in TCNs can encode both high-level and low-level features. Bokai Zhang et al. [37] propose an end-to-end temporal memory relation network for relating long-range and multi-scale temporal patterns in surgical videos.

C. Multi-branch framework

The sequential framework has been introduced into surgical phase segmentation to support accurate operating guidance. Generally, a multi-branch framework [3] deploys a refinement stage over the prediction results to perform a further refinement. Yazan Farha et al. [11] first propose a multi-branch architecture for surgical phase segmentation. For instance, they use a pre-trained I3D feature extractor and follow a causal TCN to output initial predicted results. Meanwhile, Shijie Li et al. [20] propose a multi-stage predicted structure, which consists of multiple layers of TCNs. Yueming Jin et al. [15] develops a complicated multi-task recurrent convolutional network, with a double branch framework where CNN is for tool recognition and RNN is for phase recognition. Extensive experiments have demonstrated the accuracy of surgical phase segmentation of multi-branch frameworks is higher than single-stage methods.

D. Multimodal-based approach

Considering that the multimodal-based methods are relatively sparse in recent years, we will uniformly review optical flow-related methods which apply in the surgical scene. Max Allen et al. [1]first presents a novel method for estimating the 3D pose of robotic instruments, including axial rotation, by fusing information from large homogeneous video features and local optical flow features. Gábor Lajkó et al. [19] only use of optical flow as an input for Robot-Assisted Minimally Invasive Surgery skill assessment, which shows the potential effect in surgical video training. Markus Philipp et al. [25] analyze the impact of video properties such as reflections, motion, and blur on the quality of optical flow data, and design a specifically tailored synthetic training dataset to customize the pre-trained OF-CNN for surgical activity localization. SWNet [36] is proposed with RGB stream and optical flow stream as the input data, and then encodes both of them into I3D features,

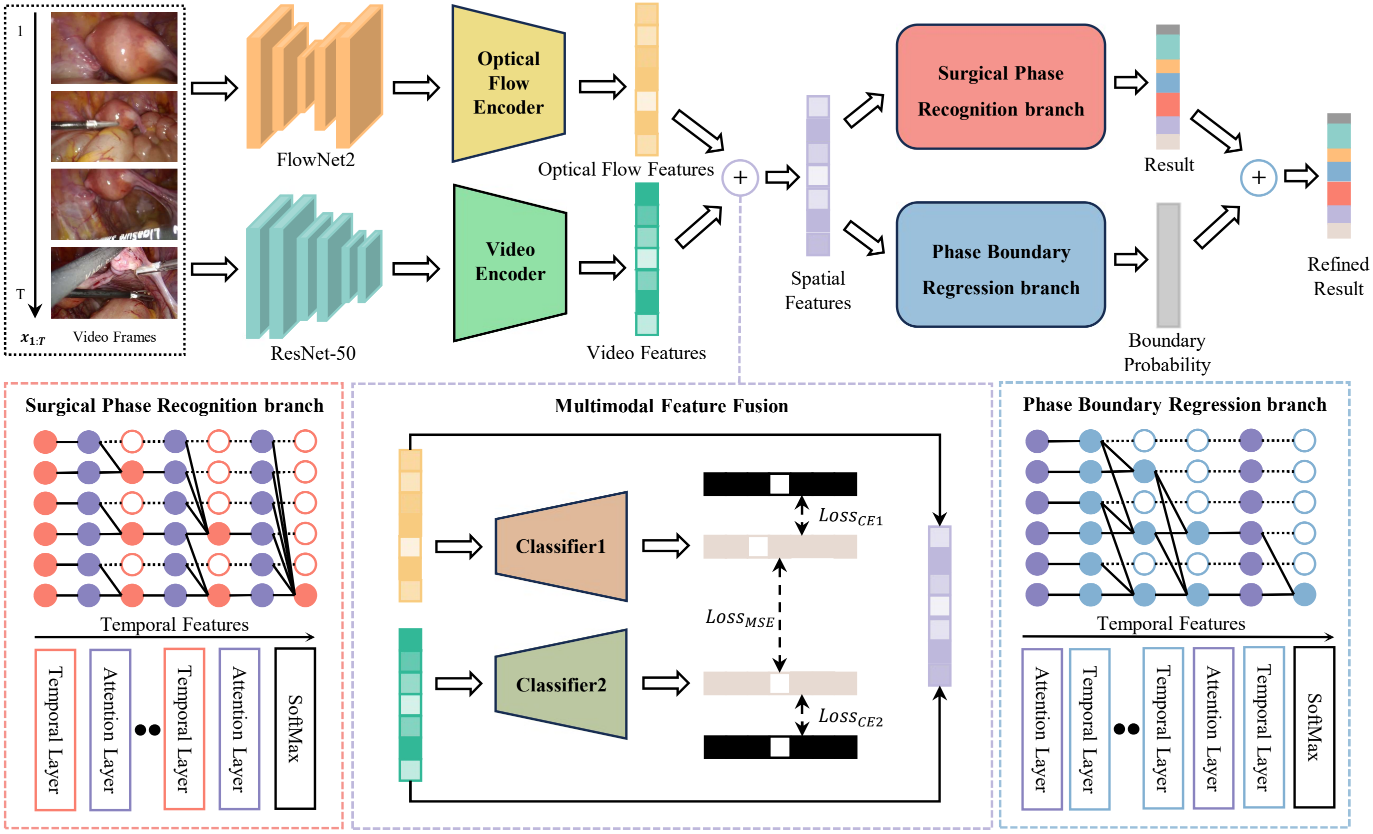


Fig. 1. Schematic illustration of proposed Surgical PHAse SEgmentation framework (SPHASE), which consists of three components: (1) Multimodal Feature Fusion (MMF) is training with triplet loss function and two fusion methods. (2) Surgical Phase Recognition branch (SPR) aims to extract the long-term dependencies from each frames. (3) Phase Boundary Regression branch (PBR) represents the short-term features.

followed by an Interaction-Preserved Channel-Separated Convolutional Network for surgical workflow recognition.

III. Method

In this section, we propose a novel Surgical Phase Segmentation Framework (SPHASE) for surgical video segmentation training tasks. As shown in Fig. 1, the overall architecture contains three components: (1) the Multimodal Feature Fusion process (MMF); (2) the Surgical Phase Recognition branch (SPR); and (3) the Phase Boundary Regression branch (PBR). Particularly, MMF process aggregates optical flow feature and I3D feature extracted from the origin video frame. In the SPR branch, we design an efficient variant of the temporal convolutional network along with a dilated attention layer to reserve more fine-grained information. In the PBR branch, we change the way to stack temporal convolutional networks with attention layers, aiming to predict surgical step boundary probability and enhance short-term dependency. Last, we refine the results from the SPR branch and PBR branch as the final surgical segmentation result.

A. Multimodal Feature Fusion process

Giving a series of surgical video frame inputs, we extract I3D features by a pre-trained ResNet-50. Meanwhile, we also extract the optical flow feature by fine-tuning FlowNet from a pair of video frames. Following two different encoders based on VisionTransformer and output two initial multimodal features F^1, F^2, denote as optical flow features and I3D features respectively.

In the surgical phase segmentation task, the optical flow pipeline and I3D pipeline theoretically have the same ground-truth labels, because they are both extracted by the same frame. Therefore, the surgical segmentation task can be regarded as a multiclass classification issue. Formally, the classification issue is as follows:

$$\widehat{p_i^1} = s(c1(\widehat{F_i^1})), \quad \widehat{p_i^2} = s(c2(\widehat{F_i^2})), \tag{1}$$

where $c1$ and $c2$ are the classifiers of optical flow features and I3D features respectively, s is the SoftMax function. Notice that they are only composed of one fully connected layer, and the SoftMax function maps the classification results between 0 and 1.

We define the multiclass classification loss, L_{CE1} and L_{CE2}, as the cross-entropy between predicted labels and ground truth y_i^{gt}:

$$\begin{aligned} L_{CE1} &= -\frac{1}{n}\sum_{i=1}^{n}[y_i^{gt}\log\widehat{p_i^1}], \\ L_{CE2} &= -\frac{1}{n}\sum_{i=1}^{n}[y_i^{gt}\log\widehat{p_i^2}], \end{aligned} \tag{2}$$

where p_i denotes the probability of the prediction is i. However, the best uni-modal performance doesn't support the best multimodal performance due to the overfitting or underfitting in the training process or suboptimal fusion method.

To avoid loss during the multimodal feature fusion process, we introduce a contrastive learning task to adjust the encoder output consistent representations across different modalities. Specifically, we define a contrastive loss shown below to make the optical flow encoding and the I3D encoding more concurrent in feature space.

$$L_{MSE} = -\frac{1}{n}\sum_{i=1}^{n} |\widehat{p_i^1} - \widehat{p_i^2}|^2. \tag{3}$$

Considering that the I3D contains more information than OF in the surgical phase segmentation task, and to avoid OF feature interfering with I3D feature, the loss L_{MSE} is only used to update the optical flow encoder.

Extracting two as close as possible features, we utilize the concatenation process as our fusion method. In general, the total loss L_{fusion} is a weighted sum of $L_{CE1}, L_{CE2}, andL_{MSE}$:

$$L_{fusion} = L_{CE1} + L_{CE2} + \alpha L_{MSE}, \tag{4}$$

where α stands for adjustable weight.

B. Surgical Phase Recognition branch

After Multimodal Feature Fusion process, we obtain the spatial feature from video frame, which aggregate by OF features and I3D features. In the SPR branch, our task is to extract the temporal features for whole video and predict the frame-wise surgical phase class C.

We observe high layers in TCN lose fine-grained information while surgical video has less inter-phase variance but more intra-phase variance. The miss fine-grained information affect modelling long-term dependencies which is more serious in surgical step recognition. To solve the problem, we introduce self-attention layer to spontaneously learn the attention weights of features, which are extracted by dilated temporal convolution network. We set a layer of residual dilated temporal convolution to extract information of high temporal reception fields, and following an attention layer to focus on learning the attention weights of features which is missing in high layers TCN. In summary, our SPR branch seems like cross-stacking several residual attention layers into a group of residual temporal networks. We connect several attention and TCN groups with a 1×1 convolution, that aiming to aggregate the most relevant information for frame-wise surgical step recognition.

Self-attention layers can be applied with a single input, which can learn weights and keep rest data static. The residual attention layer is implemented by scaled dot product operation:

$$\text{Self-Attention}(F_i) = W_q(F_i) * W_k(F_i) * W_v(F_i) \Big/ \sqrt{d}, \tag{5}$$

where W_q, W_k and W_v respectively represent three different kinds of linear transformation matrices as Query, Key and

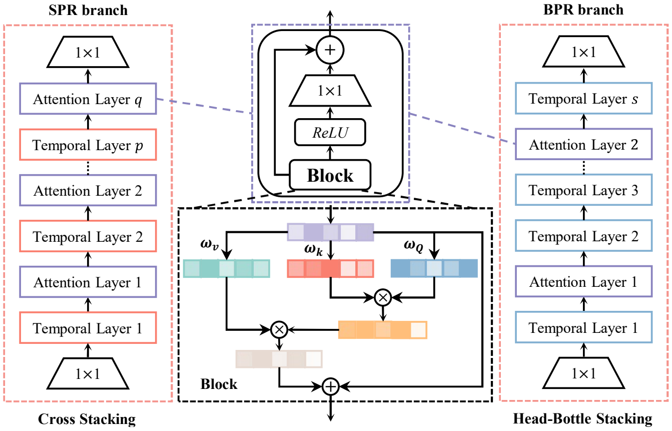


Fig. 2. Schematic illustration of Surgical Phase Recognition branch (left side), attention layer (middle side), and Phase Boundary Regression branch (right side). In SPR branch, we stack attention layer behind temporal layer, note that we don't attach attention in low layers. In PBR branch, we set attention layer next to the first temporal layer and in front of the last temporal layer. In attention block, we calculate three dimension weight numbers to gain the middle-layer features.

Value (shown in middle-bottom of Fig. 2). $1/\sqrt{d}$ is a scale factor to avoid large value of the inner product. $\sqrt{d}$ represents the largest feature dimension of those three matrices:

$$d = \text{softmax}(W_q * (W_k)^T) * W_v. \tag{6}$$

For these three linear transformation matrices, our goal is to predict what surgical step is happened, which eventually is a classification task. We use gradient descent to optimize a cross-entropy loss function as a phase classification loss:

$$L_{cls} = \sum_{t=0}^{T} -\log p_c, \tag{7}$$

where L_{cls} is the phase category probability for class c at the frame t.

To penalize over-segmentation errors, we design a combination loss function with a global loss and a local loss. The global loss aims to help the feature miner capture informative features. We apply a semantic loss to maximize the similarity between inner-phase:

$$\begin{cases} L_{local} = \frac{L_{sim}+\sigma}{L_{dif}}, \\ L_{sim} = \frac{1}{T}(\sum_{t_1,t_2 \in T} \|p_{t_1} - p_{t_2}\|), \\ L_{dif} = \frac{1}{T}(\sum_{t_1,t_2 \in T, t_1 \neq t_2 \cap p_{t_1} \neq p_{t_2}} \|p_{t_1} - p_{t_2}\|), \end{cases} \tag{8}$$

where p_t represents the phase predict result at frame t. σ is a regularizer to prevent trivial solutions, cause there are two different resolutions in Cholec80. For L_{sim}, we compute similarity of all frames to constrain the long-term features. For L_{dif}, we calculate difference between t frame with others.

We further add a global classification loss to refine features of surgical frames cause it still need to be attach to surgical phase. For each surgical frame instance P with n frames, we define:

$$L_{global} = NLL(\frac{1}{n}\sum_{t=1}^{T}(W * f_t), l_P), \tag{9}$$

where W is the weight of a classifier, and l_P is the label of predict instance P. Next, we add Gaussian Similarity-weighted TMSE (GS-TMSE) to our loss function. TMSE penalizes all frames to smooth the transition of surgical step probabilities between frames. We apply the Gaussian kernel to TMSE as follows:

$$L_{GS\text{-}TMSE} = \frac{1}{TN}\sum_{t \in T, c \in C} exp(-\frac{\|p_t - p_{t-1}\|^2}{2\sigma^2})\hat{\triangle}_{t,c}^2 \tag{10}$$

where p_t is an index of similarity for frame t and σ denotes variance. We use this function to penalize adjacent frames with large dissimilarities with a smaller weight.

In summary, the loss function for SPR is defined:

$$L_{SPR} = L_{cls} + L_{local} + L_{global} + L_{GS\text{-}TMSE}, \tag{11}$$

C. Phase Boundary Regression branch

The SPR branch aims to extract the long-term dependencies for inter-phase temporal relationship in each frames. However, we notice that over expanding receptive fields will lead to over segmentation errors. Especially for surgical video, the short-term features are also important to reserve intra-phase dependencies. In the PBR branch, our task is to predict the surgical phase boundary probability through extracting the short-term features, which is used to refine the results of surgical phase recognition. As shown in Fig. 2, differ from the way we stack attention layers and temporal layers in the SPR branch, we only add tow attention layer, one is behind the first temporal layer and other one is in front of last temporal layer. In this way, we have appropriately slowed down the expansion of receptive field and keep intra-phase consistency. PBR outputs a vector with surgical step probability $P \in [0, 1]$, which will be used for next Refinement Process to avoid over segmentation errors.

In the PBR branch, our target is to constrain the temporal transmission during intra-phase. We use a binary logistic regression loss function for minimizing intra-phase:

$$L_{PBR} = \frac{1}{T}\sum_{t=1}^{T}(s \cdot g_t \cdot \log p_t + (1 - g_t) \cdot \log(1 - p_t)), \tag{12}$$

where g_t is the ground-truth for frame t, and p_t is the phase boundary probability. We introduce a scale factor s to weight positive samples since the boundary frame is much less than other frames.

In summary, our SPHASE framework contains SPR branch for long-term dependency and PBR branch for short-term consistency. Hence, the loss function is defined as:

$$L = w_1 * L_{SPR} + w_2 * L_{PBR} + L_{fusion}, \tag{13}$$

where L_SPR and L_PBR are the loss functions for SPR branch and PBR branch respectively. L_{fusion} is the loss function for Multimodal Feature Fusion process. w_1, w_2, and w_3 are the weights of these two loss function.

TABLE I
Accuracy, Precision, Recall, Jaccard Index comparison between SOTA methods on AutoLaparo, Cholec80, M2CAI16 datasets. The best performance on trade-off for each methods is highlighted in bold.

Dataset	Methods	Acc	Pre	Rec	JACC
AutoLaparo	SV-RCNet	75.6	64	59.7	47.2
	TeCNO	77.3	66.9	64.6	50.7
	TMRNet	78.2	66	61.5	49.6
	Trans-SVNet	78.3	64.2	62.1	50.7
	LoViT	81.4	**85.1**	65.9	55.9
	SW-Net	82.8	72.1	70.9	54.5
	MS-TCN++	83.3	74.9	71.9	55.1
	Ours	**83.8**	75.7	**71.3**	**57.8**
Cholec80	EndoNet	81.7±4.2	73.7±16.1	79.6±7.9	-
	PhaseNet	78.8±4.7	71.3±15.6	76.6±16.6	-
	SV-RCNet	85.3±7.3	80.7±7	83.5±7.5	-
	TeCNO	88.6±7.8	86.5±7	87.6±6.7	75.1±6.9
	TMRNet	90.1±7.6	90.3±3.3	89.5±5	79.1±5.7
	Trans-SVNet	90.3±7.1	90.7±5	88.8±7.4	79.3±6.6
	ARST	88.46±6.81	84.93±7.83	85.05±7.24	73.16±10.17
	UATD	91.9±5.6	89.5±4.4	90.5±5.9	79.9±8.5
	LoViT	92.4±6.3	89.9±6.1	90.6±4..4	81.2±9.1
	SF-TMN	95.43±3.98	92.4±5.31	93.41±4.41	86.14±6.61
	SW-Net	93.67±3.8	91.38±1.9	92.48±8.7	78.21±9.7
	MS-TCN++	**95.76±5.82**	91.05±3.47	91.09±5.51	79.89±6.57
	Ours	94.46±7.58	**93.67±5.56**	**93.87±3.44**	**88.73±8.26**
M2CAI16	SV-RCNet	81.7±8.1	81.0±8.3	81.6±7.2	65.4±8.9
	TMRNet	87.0±8.6	87.8±6.9	88.4±5.3	75.1±6.9
	Trans-SVNet	87.2±9.3	88.0±6.7	87.5±5.5	74.7±7.7
	UATD	87.6±8.7	88.2±7.4	87.9±9.6	75.7±9.5
	SW-Net	89.0±3.5	88.6±5.3	87.6±6.7	75.9±7.1
	MS-TCN++	**91.1±3.3**	88.6±6.8	**88.5±4.1**	76.8±5.8
	Ours	90.1±7.7	**88.6±7.6**	87.6±4.2	**78.4±7.2**

In refinement process, we refine the phase recognition results R from SPR branch using phase boundary probability P from PBR branch to output the final surgical phase segmentation results. Specifically, while an element value in list P is a global maximum or over a dynamic fine-tune threshold, we mark the corresponding position at list R as a surgical phase boundary. We notice that even the most fast surgical operation is lasting for one second, so we make the frame-wise change category as among adjacent surgical phase class C. Setting each surgical phase boundary contains only one surgical step, we assign the phase name same to pre-frame surgical step class. The refinement process is only conducted during model inference.

IV. Experiment

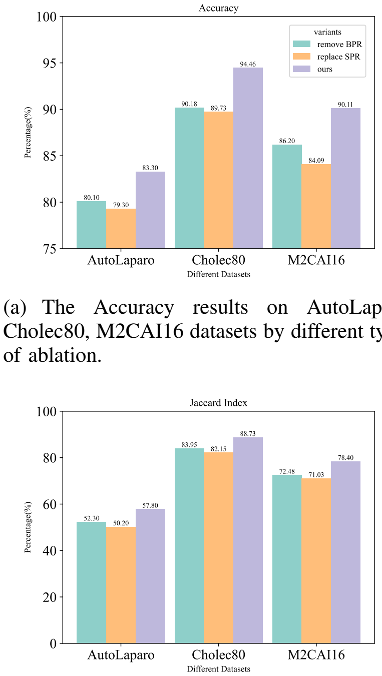
In this section, we describe the information on public experimental datasets and our evaluation metrics. Secondly, we describe the specific dataset partitioning rules and parameter settings, as well as the software environment and hardware equipment. Then we conduct extensive comparison experiments with state-of-the-art methods and analyze the results. Finally, we develop an ablation study to figure out the function of each component in SPHASE.

A. Datasets

The AutoLaparo dataset [33] contains 21 videos of laparoscopic hysterectomy. The surgical videos are recorded at 25 FPS with a standard resolution of 1920×1080 pixels. We keep

TABLE II
Ablation study on multimodal feature input. “I” and “OF” denote I3D and optical flow modalities, respectively. The best performance on Accuracy, Precision, Recall, Jaccard Index is highlighted in bold

Dataset	Method	I	OF	Acc	Pre	Rec	JACC
AutoLaparo	MS-TCN++	✓		83.3	74.9	71.9	55.1
		✓	✓	83.5	75.3	**72.5**	56.1
	SPHASE	✓		82.1	74.3	70.8	55
		✓	✓	**83.8**	**75.7**	71.3	**57.8**
Cholec	MS-TCN++	✓		95.76±5.82	91.05±3.47	91.09±5.51	79.89±6.57
		✓	✓	95.77±4.71	92.07±5.24	91.66±3.14	82.18±6.92
	SF-TMN	✓		95.43±3.98	92.4±5.31	93.41±4.41	86.14±6.61
		✓	✓	**95.84±7.89**	93.36±1.52	93.49±7.30	87.11±9.08
	SPHASE	✓		94.12±8.2	92.35±2.34	92.55±7.38	86.29±1.68
		✓	✓	94.46±7.58	**93.67±5.56**	**93.87±3.44**	**88.73±8.26**
M2CAI16	MS-TCN++	✓		91.1±3.3	88.6±6.8	88.5±4.1	76.8±5.8
		✓	✓	**91.4±6.7**	**88.7±4.4**	**88.6±5.3**	77.9±1.7
	SPHASE	✓		89.4±2.7	87.8±3.54	87.2±8.9	76.9±5.4
		✓	✓	90.1±7.7	88.6±7.6	87.6±4.2	**78.4±7.2**



(a) The Accuracy results on AutoLaparo, Cholec80, M2CAI16 datasets by different types of ablation.

(b) The Jaccard Index results on AutoLaparo, Cholec80, M2CAI16 datasets by different types of ablation.

Fig. 3. Ablation study on Multi-Branch prediction.

intact the splitting of the dataset into 10 videos for training, 4 videos for validation, and 7 videos for testing.

The Cholec80 [30] dataset includes 80 surgical videos with two different resolutions 1920*1080 and 854*480, and all videos are 25 FPS. We split the dataset into 40 videos for training, 20 videos for validation, and the remaining 20 videos for testing.

The M2CAI16 dataset [4] [28] contains 41 laparoscopic videos that are acquired at 25 FPS of cholecystectomy procedures. We use 27 videos for training, 7 videos for validation, and 7 videos for testing.

For validating our performance, we set $Accuracy$ (Acc), $Precision$ (Pre), $Recall$ (Rec), and $JaccardIndex$ ($JACC$) as the evaluation metrics in all experiments.

B. Implementation Detail

Our method is implemented with PyTorch [24] 1.13.0 + CUDA10.3. All experiments are carried out on an Intel Xeon Silver 4310 CPU (2.1 GHz), and two NVIDIA RTX 4090 GPUs. We follow the initial parameter on FlowNet [13] and pre-train the ResNet50 in ImageNet [6]. We train the I3D feature encoder with Adam for 40 epochs, with an 8-epoch warmup and cosine annealed decay. The optical flow encoder is trained for 50 epochs with Adam, the weight decay of 0.0005, the learning rate of 0.0003, and a 10-epoch warmup and cosine annealed decay. In SPR branch, the temporal convolution block consists of 10 dilated residual layers and 64 filters, with a learning rate of 0.0005, followed by a dropout layer with a dropout rate of 0.5. In PBR branch, we set temporal convolution block conclude 8 dilated residual layers and 64 filters, with a learning rate of 0.0001, batch size of 128, for a total of 30 training batches. We train the two branches separately using different loss functions, and the last refinement process is used only in the model inference phase.

C. Performance Comparison

To evaluate the effectiveness of proposed SPHASE framework, we compare it with the following 12 SOTA methods: EndoNet [30], PhaseNet [39], SV-RCNet [15], TeCNO [5], TMRNet [16], Trans-SVNet [17] [12], ARST [40], UATD [8], LoViT [22], SF-TMN [37], SW-Net [36], MS-TCN++ [20]. The corresponding numerical results are listed in Table1. We calculate the accuracy among video-level while precision, recall, and Jaccard are all in phase-level. Note that we reproduce the experiments of SW-Net and MS-TCN++ on those three surgical relevant datasets, since they provided results while got a great performance in video segmentation task. The quantitative comparison on AutoLaparo dataset is organized in the top section of Table1. Except SW-Net and MS-TCN++, the results of the other state-of-the-art methods are introduced from their respective published works. We notice that the accuracy, recall between MS-TCN++ and ours are quite small but ours is little higher. In precision metrics, LoVit is much better than other methods cause their transformer encoder. Although our method can’t catch up with LoViT in precision, it’s obviously improved in other three metrics. The quantitative comparison on Cholec80 dataset is organized in the middle section of Table I. To be mentioned that EndoNet, PhaseNet, SV-RCNet didn’t support the results of JACC. Our proposed SPHASE framework achieves the best performance on most metrics. This phenomenon is consistent across various multimodal representation types. Specifically, in JACC metric, SPHASE improve 2.6% compared with SF-TMN without precision lost. The quantitative comparison on M2CAI16 dataset is organized in the bottle section of Table1. By comparing the performance, especially the JACC, it is observed that an increase of 1.9% margin is accomplished.

TABLE III
Ablation study on different loss function. In MMF, we delete L_{CE1}, L_{CE2}, L_{MSE}, separately. In SPR branch, we remove L_{global}, $L_{GS\text{-}TMSE}$ separately and compare the results. The best performance on Accuracy, and Jaccard Index is highlighted in bold

$L_{CE1} + L_{CE2}$	L_{MSE}	L_{global}	$L_{GS\text{-}TMSE}$	L_{PBR}	Acc	$JACC$
		✓	✓	✓	72.04±8.07	65.88±7.73
	✓	✓	✓	✓	80.16±2.14	69.65±6.10
✓		✓	✓	✓	84.38±7.02	79.27±8.50
✓	✓			✓	82.40±4.61	76.85±7.46
✓	✓		✓	✓	88.74±8.42	81.71±5.41
✓	✓	✓		✓	91.49±6.62	85.11±2.71
✓	✓	✓	✓		91.64±5.32	84.76±6.05
✓	✓	✓	✓	✓	**94.46±7.58**	**88.73±8.26**

D. Ablation Study

We conducted ablation studies on the multimodal feature input, the Phase Boundary Regression branch, and the different loss function.

Multimodal Feature Input.

To demonstrate the impact of optical flow feature input, we first use only I3D feature separately, and then employ both ones on three methods: MS-TCN++, SF-TMN, and ours. Note that MS-TCN++ and SF-TMN are single feature inputting originally, therefore we re-produce the feature fusion module to finish experiments. From the results on different modality feature settings in Table II, the best performance is obtained when combine both I3D and optical flow features. This phenomenon proves the complementary of the two different feature types.

Multi-Branch Prediction.

To explore how the SPR branch and BPR branch affect the performance, we deign two variants: (1) remove the PBR branch (the framework can be operating because it only uses for outputting a refine boundary probability array); (2) replace the SPR branch to TCNs without attention layers just like TMRNet as the prediction branch. We set the accuracy and JACC as the evaluation metrics. Their performance in Figure 3 indicates that removing the PBR branch or removing the attention layers in SPR branch brings a unacceptable degradation.

Different Loss Function.

We conduct ablation study to verify the effectiveness of each part of the L. Specifically, we manually delete some factors of the loss functions in L_{SPR}, L_{PBR}, and L_{fusion} and show their performance in Table III. Note that each component of the loss function plays an irreplaceable role in our proposed framework.

E. Parameter Analysis

We conducted parameter sensitivity analysis on the hyper-parameters of main component of SPHASE framework including attention layers and temporal layers in SPR branch, temporal layers in PBR branch. Due to the limitations of submitted paper, we only list some results that are close to the best one, with focusing on accuracy and JACC. As shown in Table IV, we get the best results when the attention layer sets to 3 and temporal layer sets to 5 in SPR, and temporal layer sets to 4 in PBR, respectively.

TABLE IV
Extensive experiments for analyzing the parametric sensitivity. We set different number of attention layers in SPR branch denote as p, temporal layers in SPR as q, and temporal layers in PBR as s. Note that number of attention layers in PBR is always two. The best performance on Accuracy, Precision, Recall, Jaccard Index is highlighted in bold

p	q	s	Acc	Pre	Rec	$JACC$
2	3	5	89.30±11.40	89.52±5.59	83.43±13.42	75.31±10.65
2	5	5	90.38±7.56	89.81±6.23	782.36±11.96	4.38±11.12
3	**5**	**4**	**94.46±7.58**	**93.67±5.56**	**93.87±3.44**	**88.73±8.26**
2	6	4	90.86±6.61	90.68±5.43	81.42±12.58	73.80±11.23
3	5	6	86.71±9.67	89.84±5.01	72.74±18.66	63.97±16.09

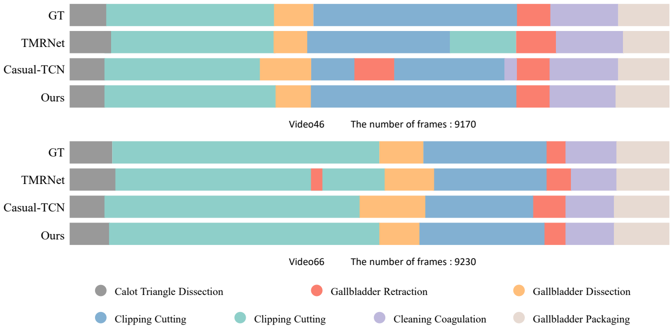


Fig. 4. Qualitative results of SPHASE on Cholec80 dataset. Compared with the ground truth, TMRNet, and Casual-TCN, we have less interruption in the inter-phase and reserve the consistency of the intra-phase.

F. Visual Results

To intuitively understand the performance improvement of SPHASE, we show some visual segmentation output of Cholec80 dataset in Figure 4. We select video 46 and 66 as samples, and our segmentation results of 98.45% and 95.37% in JACC, respectively. By comparing with the results of TMRNet and Casual-TCN, we can see that there are seldom intervals during the inter-phase in SPHASE, which demonstrates our proposed framework is able to overcome over-segmentation error and reserve the intra-phase consistency.

V. Conclusion

In this paper, we work towards surgical phase segmentation. We first propose a multimodal feature fusion mechanism for aggregating optical flow feature and I3D feature, motivated by contrast learning. We then design a surgical phase recognition branch to extract long-term dependency and a phase probability regression branch to reserve short-term consistency. Above all, we propose a new multi-feature and multi-branch surgical phase segmentation framework. Extensive experiments on AutoLaparo, Cholec80, M2CAI16 have demonstrated the effectiveness and superiority of our method.

VI. Acknowledgments

This work is being supported by Scientific and Technological Innovation 2030-Major Project of New Generation Artificial Intelligence under the Grant No. 2020AAA0109601, by the National Natural Science Foundation of China under the Grant No.62172451.

References

[1] Max Allan, Ping-Lin Chang, Sébastien Ourselin, and et al. Image based surgical instrument pose estimation with multi-class labelling and optical flow. In *Medical Image Computing and Computer Assisted Intervention*, pages 331–338. Springer International Publishing, 2015.

[2] Yutong Ban, Guy Rosman, Thomas Ward, and et al. Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows. In *2021 IEEE International Conference on Robotics and Automation*, pages 14531–14538, 2021.

[3] Binod Bhattarai, Ronast Subedi, Rebati Raman Gaire, and et al. Histogram of oriented gradients meet deep learning: A novel multi-task deep network for 2d surgical image semantic segmentation. *Medical Image Analysis*, 85:102747, 2023.

[4] Rémi Cadène, Thomas Robert, Nicolas Thome, and et al. M2CAI workflow challenge: Convolutional neural networks with time smoothing and hidden markov model for video frames classification. *CoRR*, abs/1610.05541, 2016.

[5] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, and et al. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention*, pages 343–352, 2020.

[6] Jia Deng, Wei Dong, Richard Socher, and et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] Xinpeng Ding and Xiaomeng Li. Exploring segment-level semantics for online phase recognition from surgical videos. *IEEE Transactions on Medical Imaging*, 41(11):3309–3319, 2022.

[8] Xinpeng Ding, Xinjian Yan, Zixun Wang, and et al. Less is more: Surgical phase recognition from timestamp supervision. *IEEE Transactions on Medical Imaging*, 42(6):1897–1910, 2023.

[9] Robert DiPietro, Colin Lea, Anand Malpani, and et al. Recognizing surgical activities with recurrent neural networks. In *Medical Image Computing and Computer Assisted Intervention*, pages 551–558. Springer International Publishing, 2016.

[10] Jin Fan, Ke Zhang, Yipan Huang, and et al. Parallel spatio-temporal attention-based TCN for multivariate time series prediction. *Neural Comput. Appl.*, 35(18):13109–13118, 2023.

[11] Yazan Abu Farha and Jurgen Gall. Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Xiaojie Gao, Yueming Jin, Yonghao Long, and et al. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *Medical Image Computing and Computer Assisted Intervention*, pages 593–603. Springer International Publishing, 2021.

[13] Jun Han, Jun Tao, and Chaoli Wang. Flownet: A deep learning framework for clustering and selection of streamlines and stream surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 26(4):1732–1744, 2020.

[14] Muhammad Abdullah Jamal and Omid Mohareri. Multi-modal unsupervised pre-training for surgical operating room workflow analysis. In *Medical Image Computing and Computer Assisted Intervention*, volume 13437, pages 453–463. Springer, 2022.

[15] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE Transactions on Medical Imaging*, 37(5):1114–1126, 2018.

[16] Yueming Jin, Yonghao Long, Cheng Chen, and et al. Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging*, 40(7):1911–1923, 2021.

[17] Yueming Jin, Yonghao Long, Xiaojie Gao, and et al. Trans-svnet: hybrid embedding aggregation transformer for surgical workflow analysis. *Int. J. Comput. Assist. Radiol. Surg.*, 17(12):2193–2202, 2022.

[18] Hasan Kassem, Deepak Alapatt, Pietro Mascagni, and et al. Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. *IEEE Transactions on Medical Imaging*, 42(7):1920–1931, 2023.

[19] Gábor Lajkó, Renáta Nagyné Elek, and Tamás Haidegger. Surgical skill assessment automation based on sparse optical flow data. In *2021 IEEE 25th International Conference on Intelligent Engineering Systems (INES)*, pages 000201–000208, 2021.

[20] Shijie Li, Yazan Abu Farha, Yun Liu, and et al. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6647–6658, 2023.

[21] Lei Lin, Beilei Xu, Wencheng Wu, and et al. Medical time series classification with hierarchical attention-based temporal convolutional networks: A case study of myotonic dystrophy diagnosis. In *CVPR*, pages 83–86, 2019.

[22] Yang Liu, Maxence Boels, Luis C. García-Peraza-Herrera, and et al. Lovit: Long video transformer for surgical phase recognition. *CoRR*, abs/2305.08989, 2023.

[23] Xiaoying Pan, Xuanrong Gao, Hongyu Wang, and et al. Temporal-based swin transformer network for workflow recognition of surgical video. *Int. J. Comput. Assist. Radiol. Surg.*, 18(1):139–147, 2023.

[24] Adam Paszke, Sam Gross, Francisco Massa, and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[25] Markus Philipp, Neal Bacher, Stefan Saur, and et al. From chairs to brains: Customizing optical flow for surgical activity localization. In *2022 IEEE 19th International Symposium on Biomedical Imaging*, pages 1–5, 2022.

[26] Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, and et al. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *Int. J. Comput. Assist. Radiol. Surg.*, 16(7):1111–1119, 2021.

[27] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.*, 41(4), 2022.

[28] Ralf Stauder, Daniel Ostler, Michael Kranzfelder, and et al. The TUM lapchole dataset for the M2CAI 2016 workflow challenge. *CoRR*, abs/1610.09278, 2016.

[29] Duo Tan, Jingjie Wang, Rui Yao, Jiayang Liu, Jiajing Wu, Shiyu Zhu, Ye Yang, Shanxiong Chen, and Yongmei Li. A hybrid attention mechanism based convolutional network for analysing collateral circulation via multi-phase cranial cta. In *2022 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1201–1206, 2022.

[30] Andru P. Twinanda, Sherif Shehata, Didier Mutter, and et al. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.

[31] Andru Putra Twinanda. *Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos.* PhD thesis, University of Strasbourg, France, 2017.

[32] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, and et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical Image Analysis*, 86:102770, 2023.

[33] Ziyi Wang, Bo Lu, Yonghao Long, and et al. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *Medical Image Computing and Computer Assisted Intervention*, pages 486–496. Springer Nature Switzerland, 2022.

[34] Fangqiu Yi, Yanfeng Yang, and Tingting Jiang. Not end-to-end: Explore multi-stage architecture for online surgical phase recognition. In *Computer Vision*, pages 417–432. Springer Nature Switzerland, 2023.

[35] Xin Yuan, Zhe Lin, Jason Kuen, and et al. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

[36] Bokai Zhang, Amer Ghanem, Alexander Simes, and et al. SWNet: Surgical workflow recognition with deep convolutional network. In *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143, pages 855–869. PMLR, 2021.

[37] Bokai Zhang, Mohammad Hasan Sarhan, Bharti Goel, and et al. SF-TMN: slowfast temporal modeling network for surgical phase recognition. *CoRR*, abs/2306.08859, 2023.

[38] Yitong Zhang, Sophia Bano, Ann-Sophie Page, and et al. Retrieval of surgical phase transitions using reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention*, pages 497–506. Springer Nature Switzerland, 2022.

[39] Weiqiang Zhu and Gregory C Beroza. Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 2018.

[40] Xiaoyang Zou, Wenyong Liu, Junchen Wang, and et al. ARST: autoregressive surgical transformer for phase recognition from laparoscopic videos. *CoRR*, abs/2209.01148, 2022.