SPRMamba: Surgical Phase Recognition for Endoscopic Submucosal Dissection with Mamba

Xiangning Zhang[a,1], Jinnan Chen[b,1], Qingwei Zhang[b,1], Chengfeng Zhou[a], Zhengjie Zhang[a], Xiaobo Li[b,*], Dahong Qian[a,*]

[a]*School of Biomedical Engineering, Shanghai Jiao Tong University, ShangHai, 200000, China*
[b]*Division of Gastroenterology and Hepatology, Shanghai Institute of Digestive Disease, NHC Key Laboratory of Digestive Diseases, Renji Hospital,, Shanghai Jiao tong University School of Medicine, ShangHai, 200000, China*

Abstract

Endoscopic Submucosal Dissection (ESD) is a minimally invasive procedure initially designed for the treatment of early gastric cancer but is now widely used for various gastrointestinal lesions. Computer-assisted Surgery systems have played a crucial role in improving the precision and safety of ESD procedures, however, their effectiveness is limited by the accurate recognition of surgical phases. The intricate nature of ESD, with different lesion characteristics and tissue structures, presents challenges for real-time surgical phase recognition algorithms. Existing surgical phase recognition algorithms struggle to efficiently capture temporal contexts in video-based scenarios, leading to insufficient performance. To address these issues, we propose SPRMamba, a novel Mamba-based framework for ESD surgical phase recognition. SPRMamba leverages the strengths of Mamba for long-term temporal modeling while introducing the Scaled Residual TranMamba block to enhance the capture of fine-grained details, overcoming the limitations of traditional temporal models like Temporal Convolutional Networks and Transformers. Moreover, a Temporal Sample Strategy is introduced to accelerate the processing, which is essential for real-time phase recognition in clinical settings. Extensive testing on the ESD385 dataset and the cholecystectomy

*Corresponding author.
Email addresses: zxnyyyyy@sjtu.edu.cn (Xiangning Zhang), lxb_1969@163.com (Xiaobo Li), dahong.qian@sjtu.edu.cn (Dahong Qian)
[1]The three authors contribute equally to this work.

Cholec80 dataset demonstrates that SPRMamba surpasses existing state-of-the-art methods and exhibits greater robustness across various surgical phase recognition tasks.

Keywords: Endoscopic submucosal dissection, surgical phase recognition, surgical video analysis, Mamba

---

1. Introduction

Endoscopic Submucosal Dissection (ESD) is a groundbreaking minimally invasive procedure that has transformed the management of early gastrointestinal cancers. Compared to traditional surgery, ESD offers patients reduced trauma, faster recovery times, and lower rates of cancer recurrence [1]. However, the intricacy of lesion characteristics and the delicate nature of the gastrointestinal tract pose a risk of unexpected complications during ESD, making proficiency in this technique limited to a few experienced endoscopists. With the increasing detection of gastrointestinal cancers through routine gastric endoscopy screenings, there is a growing demand for skilled endoscopists capable of performing ESD. Computer-assisted surgery (CAS) systems are an advanced medical technology that can significantly enhance surgical efficiency and reduce the risk of complications [2]. Surgical phase recognition (SPR) is a critical and challenging task within CAS systems, aimed at identifying surgical activities from surgical videos [3]. In ESD, the phases carry distinct risks, necessitating accurate phase recognition for real-time monitoring, surgical process optimization, context-aware decision support, and early anomaly detection [4]. Additionally, automated SPR can significantly improve postoperative reporting and video indexing, providing valuable educational resources for novice endoscopists [5]. Therefore, developing efficient and accurate video-based SPR algorithms is essential to meet the demands of modern surgical practice and education.

Limited inter-phase variance and high intra-phase variance are the most critical challenges for automatically recognizing surgical phases from video. Even though a lot of surgical phase recognition methods have been proposed to address this issue, for example, Jin et al. [6] introduced TMRNet, which includes a repository to store remote information to learn the relationship between the current frame and previous frames. Czenpiel et al. [7] proposed TeCNO, an enhanced Multi-Stage Temporal Convolutional Network (MS-TCN) [8]. Gao et al. [9] devised Trans-SVNet, addressing fine-grained

temporal features loss in TCNs with a compact Transformer model. However, ESD surgical phase recognition is still challenged since their surgical phases are extremely complex and temporally imbalanced. Therefore, applying existing surgical phase recognition methods directly to ESD data leads to performance degradation [10, 11, 12]. A more efficient temporal modeling approach is needed to address the insufficient temporal modeling capabilities of traditional temporal models.

Recently, a state-space model called Mamba has demonstrated substantial potential in modeling temporal contexts. Unlike traditional temporal models, the Mamba model shows higher robustness and accuracy in dealing with complex surgical phase transitions while avoiding the problem of the Transformer's high computational complexity. Making Mamba well suited for modeling the temporal context in ESD videos. To fully utilize Mamba's temporal modeling ability, we propose a Mamba-based surgical phase recognition framework called SPRMamba. SPRMamba leverages a ResNet-50 backbone for spatial feature extraction from ESD videos, followed by the application of four LSTContext modules to effectively combine short-term and long-term temporal contexts for phase recognition. Central to SPRMamba is the Scale Residual TranMamba (SRTM) module, which combines Mamba's long-term temporal modeling capabilities with the Transformer's ability to capture fine-grained details, addressing the complex phase relationships inherent in ESD surgeries. Furthermore, according to [13], allowing the model to operate on the complete input sequence is more beneficial compared to only accessing a subset of the input. Therefore, considering the quadratic complexity of the transformer and the high memory usage of $\mathcal{O}(L^2)$ (where L represents the input sequence length), it is inappropriate to directly apply the transformer to uncut surgical videos. To address this issue, we designed two sampling strategies, a Long-range Sample and a Window Sample, to reduce computational complexity and support the online phase. Our main contributions can be summarized as follows:

- We propose a novel surgical phase recognition framework (SPRMamba), leveraging the Scaled Residual TranMamba (SRTM) module to efficiently model short-term and long-term temporal contexts.

- Introducing Temporal Sample Strategy (TSS) with Window and Long-range sampling to reduce computational burden and support online phase recognition.

- Conducting qualitative and quantitative experiments on ESD385 and Cholec80 datasets, demonstrating significant improvements over existing techniques.

2. Related Work

2.1. Video understanding

In the field of video understanding, researchers are dedicated to developing models that can effectively capture spatiotemporal features to extract semantic information from dynamic scenes. Traditional video understanding methods often rely on 3D Convolutional Neural Networks (3D-CNNs) [14], which extend 2D convolutions to simultaneously process spatial and temporal dimensions. However, 3D-CNNs tend to have high computational complexity, limiting their application in long video analysis.

In the field of video understanding, effectively capturing spatiotemporal features is crucial for extracting semantic information from dynamic scenes. Early approaches, such as 3D Convolutional Neural Networks (3D-CNNs) [14], extended 2D convolutions to process spatial and temporal dimensions simultaneously, achieving reasonable performance in action segmentation tasks. However, the high computational complexity of 3D-CNNs makes them impractical for analyzing long videos, particularly in medical applications like ESD surgery, where procedures can last several hours. To address computational limitations, Temporal Convolutional Networks (TCNs) [15] emerged, which focus on capturing temporal dependencies through one-dimensional convolutions. While TCNs are more efficient and capable of handling longer video sequences, they struggle to model complex action variations and capture the fine-grained details needed for surgical phase recognition, where subtle movements and transitions between phases play a critical role. To address this, Temporal Convolutional Networks (TCNs) [15] were proposed, focusing on capturing temporal information through one-dimensional convolutions, thereby modeling long-term dependencies with lower computational costs. However, TCNs face limitations in handling complex action variations and capturing fine-grained features. Transformer-based models have gained attention in recent years for their ability to model long-range dependencies using self-attention mechanisms [16]. While Transformers achieve state-of-the-art performance in video understanding tasks by capturing spatiotemporal relationships across frames, they are constrained by their quadratic computational complexity, which poses challenges when applying them to

long-duration videos in real-time scenarios, such as ESD surgeries. Despite these advancements, the complexity of ESD surgery poses unique challenges for current methods in video understanding. Surgical phases often involve intricate and fine-grained transitions that are difficult to capture with traditional models. In addition, effectively modeling short-term and long-term spatiotemporal dependencies in long videos while maintaining controllable computational costs remains a major challenge.

2.2. Surgical Phase Recognition

Research in the field of Surgical Phase Recognition mainly focuses on automatically detecting and classifying different phases of surgery by analyzing surgical videos. Early surgical phase recognition methods relied heavily on hand-designed features and statistical learning methods [17, 18, 19, 20, 3, 21, 22, 23]. However, these methods are limited by empirical design features and usually rely on predefined dependencies, making it difficult to accurately capture subtle motion features with strongly nonlinear dynamics, and their performance suffers greatly. Twinanda et al. [3] proposed the EndoNet model, a CNN-based approach that automatically extracts features and performs phase classification from laparoscopic surgical videos. This approach opened the way for SPR research using deep learning. Since then, researchers have sought to enhance SPR by incorporating more advanced temporal modeling techniques. For instance, long short-term memory (LSTM) networks have been widely adopted in models like PhaseNet [24], EndoLSTM [25], and SV-RCNet [26], enabling the capture of both spatial and temporal dependencies. These methods, by combining deep residual networks (ResNet) with LSTM modules, have shown promising improvements in recognizing surgical phases by effectively modeling temporal series data. However, these methods are computationally expensive when dealing with long time series and prone to misclassification when dealing with complex phase transitions. To address these limitations, Transformer-based architectures have been introduced into SPR due to their capability of capturing long-range temporal dependencies. For example, Czempiel et al. [27] proposed a Transformer-based surgical phase recognition method, which significantly improves the model's performance in complex surgical scenarios through the self-attention mechanism. More recently, Graph Neural Networks (GNNs) have been employed in SPR, with studies such as Padoy et al. [28] demonstrating their effectiveness in modeling complex spatiotemporal relationships between surgical tools and anatomical structures. Although existing techniques have made significant

progress in SPR, there are still some challenges, such as the real-time nature of the model, and the accurate recognition of short-time phases. Therefore, this paper aims to construct a novel ESD surgical phase recognition model that is capable of maintaining high accuracy while reducing computational complexity.

2.3. State Space Models

Recently, State-Space Models (SSMs) have been proven to have transformer-level performance in capturing long sequences in the field of natural language processing [29, 30]. [29] introduced a new model for Structured State-Space Sequences (S4), specifically designed to model long-range dependencies exhibiting the well-established property of scaling linearly with sequence length. Based on this, [31] proposed an advanced layer called S5, which integrates MIMO SSM and efficient parallel scanning into the S4 architecture. This development aims to overcome the limitations of SSMs and improve their efficiency. In addition, [32] contributed a novel SSM layer H3, significantly narrowing the performance gap between SSM and transformer-based attention in language modeling. [33] Expand the S4 model by introducing additional gating units in the gated state space layer to enhance its expressiveness. More recently, [30] developed a universal language model called Mamba by introducing a data-dependent SSM layer and a selection mechanism using parallel scanning. Compared to transformers based on quadratic complexity attention, Mamba excels at handling long sequences with linear complexity.

In the field of vision, [34] proposed Visual Mamba (Vim), which combines position encoding and bi-directional scanning to efficiently capture the global context of an image. Pioneered the application of Mamba in vision tasks. Li et al. [35] constructed a generic framework called Video Mamba Suite to develop, validate, and analyze Mamba's performance in video understanding. Besides, the great potential of Mamba has inspired a series of works [36, 37, 38, 39, 40], which demonstrated Mamba's has better performance and higher GPU efficiency than Transformer on visual downstream tasks such as semantic segmentation and video understanding. In our work, we integrate Mamba into the surgical phase recognition model to efficiently capture the temporal context in surgical video.

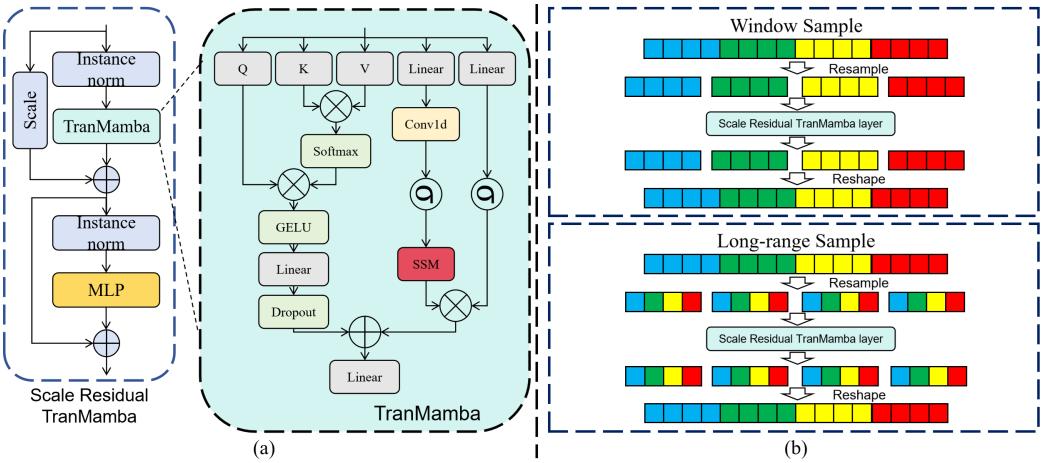

Figure 1: A schematic overview of the LSTContext block; (a) This represents the Scale Residual TranMamba (SRTM), which is composed of the TranMamba and its main components; (b) The Temporal Sample Strategy is illustrated with a window of size 4. For window sampling, the sequence is partitioned into small windows, and the SRTM is computed for each window. For long-term sampling, the sequence is reordered such that the SRTM is computed over the entire, but sparsely sampled sequence. After the SRTM computation, the output is reordered to preserve the original sequence order.

3. METHODOLOGY

3.1. Preliminaries

State Space Models (SSMs) are typically considered linear time-invariant systems that map a 1-D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^N$. These systems are commonly represented by linear ordinary differential equations (ODEs) with $\mathbf{A} \in \mathbb{R}^{N \times N}$ as the evolution parameter, and $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$ as projection parameters.

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t). \tag{1}$$

Structured State Space Sequence Models (S4) and Mamba are discrete versions of continuous systems, which include a time scale parameter $\mathbf{\Delta}$ that converts continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$. A common conversion method is Zero-Order Hold (ZOH), defined as follows:

$$\bar{\mathbf{A}} = \exp(\mathbf{\Delta A}), \bar{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}. \tag{2}$$

After the discretization of $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, Eq.(1) can be expressed as discrete parameters:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, y_t = \mathbf{C}h_t. \tag{3}$$

Finally, for an input sequence of size T, the output y is calculated using a global convolution operation with a convolution kernel $\bar{\mathbf{K}}$

$$\bar{\mathbf{K}} = (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, ..., \mathbf{C}\bar{\mathbf{A}}^{\mathbf{M-1}}\bar{\mathbf{B}}), \mathbf{y} = \mathbf{x} * \bar{\mathbf{K}}, \tag{4}$$

where M is the length of the input sequence x, and $\bar{\mathbf{K}} \in \mathbb{R}^M$ is a structured convolutional kernel.

3.2. Scale Residual TranMamba

For the surgical phase recognition task, temporal information plays a crucial role in accurate recognition. However, due to the complexity of ESD surgeries, existing surgical phase recognition algorithms based on traditional temporal models are difficult to apply directly to ESD data. We aim to explore a simple and effective structure combining Mamba that simultaneously models both long-term and short-term temporal contexts. A straightforward approach is to combine the short-term temporal context modeled by the transformer with the long-term temporal context modeled by Mamba. Therefore, we propose the Scaled Residual TranMamba module, as shown in Fig. 1a.

Given the input feature $\mathbf{F_{in}} \in \mathbb{R}^{L\times C}$, the SRTM module first applies Instance norm[41]. Then, it uses the TranMamba module to capture long-term and short-term temporal context, thereby generating $\mathbf{F_{ls}} \in \mathbb{R}^{L\times C}$. In addition, to obtain more comprehensive context information, the fusion of $\mathbf{F_{in}}$ and $\mathbf{F_{ls}}$ was achieved through scale residual connections. The fused feature is followed by normalization utilizing Instance Norm and then MLP to learn deeper features. The entire process can be described as follows:

$$\mathbf{F} = \beta\mathbf{F_{in}} + TranMamba(IN(\mathbf{F_{in}})) \tag{5}$$

$$\mathbf{F_{out}} = MLP(IN(\mathbf{F})) \tag{6}$$

Specifically, there are three branches in the TranMamba module. The first branch takes the first quarter portion of the input feature $\mathbf{F_{in}}$ along the channel dimension as input $\mathbf{F_{b1}} \in \mathbb{R}^{L\times\frac{C}{4}}$, then expands the dimension to $\lambda\mathbf{C}$ by a linear transformation, and then activates it using the SiLU[42] function. The second branch in the TranMamba module has inputs similar to the first branch. It takes the second quarter of the input features $\mathbf{F_{in}}$ along the channel dimension as input $\mathbf{F_{b2}} \in \mathbb{R}^{L\times\frac{C}{4}}$. Subsequently, these features are sequentially expanded through the dimensional expansion of the

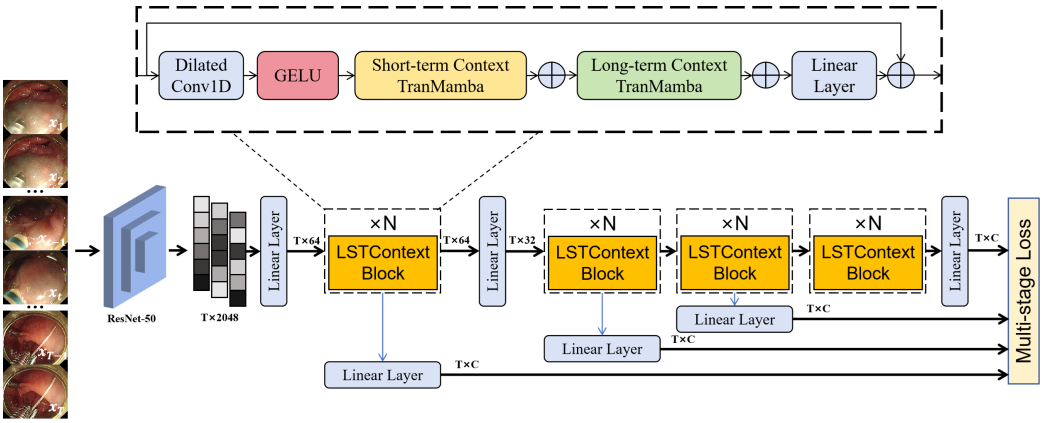

Figure 2: A schematic overview of the proposed SPRMamba architecture, which consists of a ResNet-50, and four LSTContext blocks (top).

linear layer, Conv1d layer, SSM, and LayerNorm. Afterward, the features extracted from both branches are fused using the Hadamard product, which is designed to capture the long-term temporal context information in this way[30]. The third branch takes the last half of the input features $\mathbf{F_{in}}$ along the channel dimension as input $\mathbf{F_{b3}} \in \mathbb{R}^{L \times \frac{C}{2}}$. Subsequently, $\mathbf{F_{b3}}$ is input into the Attention layer and activated using the GELU function[43], aiming to capture short-term temporal context features. Finally, the short-term temporal features achieved from the third branch will be fused with the previously captured long-term temporal features via a linear mapping to obtain feature $\mathbf{F_{ls}}$. The entire process takes the following form:

$$\mathbf{F_1} = SiLU(Linear(\mathbf{F_{b1}})) \tag{7}$$

$$\mathbf{F_2} = LN(SSM(Conv1d(Linear(\mathbf{F_{b2}})))) \tag{8}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{K}}V) \tag{9}$$

$$\mathbf{F_3} = Dropout(Linear(GELU(Attention(\mathbf{F_{b3}})))) \tag{10}$$

$$\mathbf{F_{ls}} = Linear(\mathbf{F_1} \odot \mathbf{F_2} + \mathbf{F_3}) \tag{11}$$

where $\odot$ represents Hadamard product, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{\mathbf{L} \times \mathbf{D}}$ are linearly transformed form $\mathbf{F_{b3}}$

3.3. Temporal Sample Strategy

Due to the quadratic complexity of self-attention blocks, it is impractical to apply attention to long sequences of untrimmed videos. This is because the L of the video sequence is very large, so we need to resample it to achieve modeling of long-term and local temporal contexts, as shown in Fig. 1b.

Temporal Window Sample For temporal window sampling (WS), we divide the sequence into non-overlapping windows of size W. Fig. 1b illustrates the case of W=4. Given that different tasks may have different time-dependent ranges, W is task-specific. We used W=64 in practice. The impact of W was evaluated in Section 4. Instead of modeling the temporal context over the entire sequence of length T, we model the temporal context $\frac{\mathbf{T}}{\mathbf{W}}$ times, where each $\mathbf{F_{b1}} \in \mathbb{R}^{W \times \frac{C}{4}}$, $\mathbf{F_{b2}} \in \mathbb{R}^{W \times \frac{C}{4}}$, $\mathbf{F_{b3}} \in \mathbb{R}^{W \times \frac{C}{2}}$ corresponds to each window.

Temporal Long-range Sample For the temporal long-range sample, we sample the input every G to divide G non-overlapping sequences and model the temporal context information of each sequence. In $\mathbf{G = 4}$ shown in Fig. 1 b, we model the temporal context information of each of the four downsampled sequences. In general, we model the temporal contexts for G $\mathbf{F_{b3}} \in \mathbb{R}^{W \times \frac{C}{2}}$ where the $\mathbf{F_{b1}}$ and $\mathbf{F_{b2}}$ are the same, *i.e.*, $\mathbf{F_{b1}}, \mathbf{F_{b2}} \in \mathbb{R}^{W \times \frac{C}{4}}$. The parameter G provides the flexibility to adjust the sparsity based on the available memory budget, e.g., G=1 corresponds to the case where the attention is applied to the entire sequence. In practice, we use G=64 and evaluate the impact of G in Section 4.

LSTContext Block The top of Fig. 2 illustrate the entire LSTContext block. As in previous work [7], we use a 1D dilated convolution with kernel size 3. This is because the dilated convolution increases the receptive field without the need to increase the parameter number by increasing the kernel size. Where the dilation factor for each layer increases by a factor of 2 and the receptive field exponentially expands as the number of layers increases. Therefore, with a few parameters, we achieved a significantly large receptive field in the temporal sequence, which mitigated model overfitting and effectively promoted the accuracy of surgical phase recognition. The dilated convolution is followed by a Gaussian Error Linear Unit (GELU). In the LSTContext block, we first use the window sample and then the long-range sample, as shown in Fig. 1. Finally, we use a linear layer with residual connectivity to output the features for each frame, $\mathbf{F} \in \mathbb{R}^{L \times D}$

3.4. Overview

The architecture of the proposed SPRMamba is shown in Fig. 2. For a video $\mathbf{V} \in \mathbb{R}^{L\times C\times H\times W}$ of length L, we first extract the spatial frame-level feature sequence $\mathbf{F} \in \mathbb{R}^{L\times D}$ from a fixed ResNet-50[44], where D=2048 is the spatial dimension, and then SPHMamba uses a linear layer to reduce the feature dimension to 64. As in previous works, we repeated each LST-Context block N times, where the dilation factor of the dilation convolution was increased in each layer. After the first N layers of the LTContext block, we use an additional linear layer to further reduce the dimensionality D to 32. The dimensionality reduction method reduces the number of parameters from 2.49 million to 1.23 million without degrading the accuracy. We also use an additional Conv1D followed by a softmax layer to generate the frame-level class probabilities $\mathbf{P} \in \mathbb{R}^{L\times C}$. We proceed with three additional stages, each consisting of N layers of LSTContext blocks. At the beginning of each stage, we reset the dilation factor of the temporal convolution to 1 and compute the frame-level class probabilities $P \in \mathbb{R}^{T\times C}$ after each stage, resulting in a multi-stage loss. We use cross-entropy loss and the mean squared error smoothing loss introduced by [8] for supervised training.

4. Experiments

In this section, we first describe the dataset used in our study, as well as the detailed experimental setup and evaluation metrics employed. Subsequently, we validated the effectiveness of our proposed method in ESD and cholecystectomy by comparing it with the SOTA method and conducting ablation experiments.

4.1. Dataset

Table 1: Distribution of annotations in the ESD385 dataset for phase recognition.

ESD Surgical phases	Preparation	Estimation	Marking	Injection	Incision	ESD	Vessel-treatment	Clips	Total
Train	91739	43198	15585	44464	59695	174201	44317	10817	484016
val	14193	5822	2820	9886	11917	41807	4865	3985	95295
test	14323	16860	6583	15831	22181	75968	14707	4748	171201

4.1.1. ESD385

There is no publicly available dataset on ESD surgery. To validate the effectiveness of the proposed algorithm in this study, we retrospectively selected

patients who underwent ESD from August 16, 2023, to January 8, 2024, in the endoscopy unit of the Department of Gastroenterology, Renji Hospital, Shanghai Jiao Tong University. A total of 385 videos of ESD procedures were collected. All procedures were performed by experienced endoscopists. The instruments used for ESD procedures included: gastroscope GIF-Q260, GIF-H260, GIF-HQ290, mucosal incision knife KD-612L/U, KD-655L/U, KD-620UR, endoscopic water pump OFP-2, mucosal injection needle NM-400L-0423, NM-400U-0423, thermal biopsy forceps HDBF-2.4-230-S, hemostatic forceps FD-410LR, FD-412LR, soft tissue clips ROCC-D-26-195-C, ROCC-D-26-235-C, hemostatic clips M00521242, methylene blue injection, indigo rouge solution. All endoscopic surgical videos were recorded using Image Management Hub, IMH-200, and Olympus and stored in MP4 format. The video resolution was 1920 × 1080 with a frame rate of 50 fps. After collection, all videos were anonymized to remove identifiable patient information, procedure timestamps, and other identifying details. This work was approved by the Renji Hospital affiliated with the Shanghai Jiao Tong University School of Medicine ethics committee with number RA-2024-457(2024.3.20).

The ESD video is divided into eight surgical phases, including (1) Estimation; (2) Marking; (3) Injection; (4) Incision; (5) ESD; (6) Vessel-treatment; (7) Clips(portion); and (8) Preparation. Phase (6) Preparation included elements unrelated to the ESD procedure, such as gastric insufflation and device replacement. Four endoscopists independently annotated the video, labeling each frame in the video. After the initial annotation, quality control was performed by two additional experienced endoscopists. Any uncertainties that arose during the quality control process were resolved through collaborative discussions among the six experts. The number of annotated frames in the dataset varied at each phase, with the ESD phase occupying most of the procedure time, which is the most important and skill-demanding phase of ESD. Detailed statistics for each phase are shown in Table 7. Overall, a total of 484016 frames, 95,295 and 171,201 frames were annotated for training, validation, and verification, respectively.

4.1.2. Cholec80

Furthermore, to verify the robustness of our proposed algorithm, we validated it on the Cholec80 dataset[3] for cholecystectomy procedures. The Cholec80 dataset is a large-scale surgical benchmark dataset containing 80 cholecystectomy videos performed by 13 surgeons. These videos are recorded at 25 frames per second, with each frame having a resolution of either

1920×1080 or 854×480 pixels. Each frame in the videos is annotated with one of seven surgical phases: Preparation (P1), Calot triangle dissection (P2), Clipping and cutting (P3), Gallbladder dissection (P4), Gallbladder packaging (P5), Cleaning and coagulation (P6), and Gallbladder retraction (P7). Additionally, each frame includes annotations for seven types of surgical tools, including Grasper, Bipolar, Hook, Clipper, Scissors, Irrigator, and Specimen bag. We strictly followed the experimental protocol used in previous studies[3, 26, 6], dividing the dataset into two equal subsets: the first 40 videos for training and the remaining 40 videos for testing, with 8 videos from the training set selected for validation.

4.2. Evaluation Metrics

For surgical phase recognition, we employed four commonly used evaluation metrics to quantitatively assess model performance, which have been used in previous phase recognition work as well[3, 26, 45, 6]. These metrics include Precision, Recall, Jaccard Index, and Accuracy. Accuracy is defined as the percentage of frames across the entire video correctly predicted to be in their ground truth phase. Given the imbalanced phase presented in the video, Precision, Recall, and Jaccard Index refer to phase-level evaluations, calculated within each phase and then averaged across all phases.

4.3. Implementation details

Our approach is implemented in Python using Pytorch framework and training is conducted on a workstation equipped with an NVIDIA RTX 4090. In the first stage, we use a ResNet-50 model pre-trained on ImageNet-22K[44]. We then fine-tune the model on our data. To ensure a fair comparison with SOTA methods, we downsample all videos to 1 fps, which is also the approach used in previous works[45, 6]. This operation has additional benefits, including enriching temporal information and saving memory. During training, image frames are further downsampled from the original resolutions of 1920 × 1080 and 854 × 480 to a resolution of 250 × 250 pixels to further reduce memory usage. Data augmentation is performed through 224 × 224 cropping, flipping, and random mirroring to expand the training dataset. The ResNet-50 model is fine-tuned with a batch size of 160 images. In the second stage, LSTContext is trained end-to-end from scratch. The W and G for each task in the LSTContext module are initially chosen based on estimates of task average durations on the training dataset and are then fine-tuned through ablation studies. For both stages, we adopted the same

experimental setup as the literature[44], using the AdamW optimizer[46] with a weight decay of 1e-5 and approximately 200 iterations. A cosine learning rate scheduler[47] is used, with 40 epochs of linear warm-up and an initial learning rate of 5e-4.

4.4. Comparison with State-of-the-Art Methods

4.4.1. Result on the ESD385 Dataset

Table 2: COMPARISON WITH THE SOTA METHODS ON THE ESD385 DATASET.

Method	Accuracy(%)	Precision(%)	Recall(%)	Jaccard(%)	FLOPs	Params
ResNet-50 [44]	72.31	78.75	66.63	55.39	4.1G	24.56M
SV-RCNet [26]	75.58±13.46	81.27±17.23	70.78±17.46	60.12±17.77	41.1G	28.76M
SAHC [48]	86.64±10.63	86.35±18.07	83.75±17.55	75.14±18.20	9.5G	26.27M
Furube et at. [12]	84.79±11.58	84.85±18.75	82.11±17.52	72.60±18.25	4.5G	24.69M
AI-Endo [11]	83.38±12.09	84.74±17.35	82.17±17.23	72.09±17.23	5.7G	24.72M
SPRMamba(Ours)	**87.64±9.83**	**86.72±16.54**	**86.76±15.66**	**77.51±17.83**	7.5G	25.42M

On the ESD385 dataset, we compared our proposed method with the SOTA method. This includes the method proposed by Furube et al. [12], which fine-tunes ResNet-50 as a feature extractor and then uses MS-TCN for hierarchical prediction refinement for surgical phase recognition. The method proposed by Cao et al. [11]. An Intelligent Surgical Workflow Recognition Suite for ESD is based on Trans-SVNet [9] implementation. In addition, we compared our method with SV-RCNet [26] and SAHC [48], two SOTA methods for cholecystectomy surgical phase recognition. The comparative results are presented in Table 2. We found that applying existing surgical phase recognition methods directly to ESD leads to performance degradation because ESD surgery has more complex workflows compared to traditional surgeries. In the case of image classification using ResNet-50 alone, the average accuracy is 72.31%, the average precision is 78.75%, the average recall is 66.63%, and the average Jaccard is 55.39%. After including the SRTM module and Temporal Sample Strategy, the present method achieves an average accuracy of 87.64%, an average precision of 86.72%, an average recall of 86.76%, and an average Jaccard of 77.51%, increasing the average accuracy from 72.31% to 87.64%. These results demonstrate the effectiveness of the method for automatic phase recognition. Compared to the SOTA method, our method outperforms the second-best method in all four metrics by 1.00%, 0.37%, 3.01%, and 2.37%, respectively. Furthermore, to demonstrate the algorithm efficiency of the proposed method. We also compare the number of parameters and computational cost of our method with the

SOTA method, which is reported in Table 2. "Params" denotes the number of parameters. Our proposed method outperforms SAHC in terms of fewer parameters. Compared to SAHC, our model will reduce 2.0G FLOPS.

In Fig. 3, we visualized the predictions of two ESD videos to qualitatively show the improvement in surgical phase recognition. The results highlight that SPRMamba obtains consistent and smooth predictions not only within one phase but also for the often ambiguous phase transitions. Compared with AI-Endo and Furube et at., SPRMamba can perform accurate phase recognition even for phases with short durations, such as phase Preparation and Clips. Finally, SPRMamba shows robustness because Video 2 lacks phase Clips. However, the performance of our model does not deteriorate.

4.4.2. Result on the Cholec80 Dataset

Table 3: COMPARISON WITH THE SOTA METHODS ON THE Cholec80 DATASET. Note that the '+' denotes methods based on multi-task learning that require extra tool labels.

Method	Accuracy(%)	Precision(%)	Recall(%)	Jaccard(%)
Endonet [3]$^{+}$	81.70±4.20	73.70±16.10	79.60±7.90	-
MTRCNet-CL [45]$^{+}$	89.20±7.60	86.90±4.30	88.00±6.90	-
LAST [49]$^{+}$	93.12±4.71	89.25±5.49	90.10±5.45	81.11±7.62
SV-RCNet [26]	85.30±7.30	80.70±7.00	83.50±7.50	-
TeCNO [7]	89.35±6.70	83.24±7.21	81.29±6.61	70.08±9.08
TMRNet [6]	90.10±7.60	**90.30±3.30**	89.50±5.00	79.10±5.70
Trans-SVNet [9]	90.27±6.48	85.23±6.97	82.92±6.77	72.42±8.92
SAHC [48]	91.80±8.10	90.30±6.40	90.00±6.40	81.20±5.50
SPRMamba(Ours)	**93.12±4.58**	89.26±6.69	**90.12±5.61**	**81.43±6.90**

In the Cholec80 dataset, we conducted a performance comparison of our methods with state-of-the-art (SOTA) methods, including Endonet [3], MTRCNet-CL [45], LAST [49], SV-RCNet [26], TeCNO [7], TMRNet [6], Trans-SVNet [9], and SAHC [48]. Please note that we re-implemented TeCNO and Trans-SVNet using the model weights provided in the original manuscript. The results of the other state-of-the-art methods were extracted verbatim from their respective published works. Our comparison results are presented in Table 3. Our method outperforms the other methods in most of the evaluation metrics, except for the average precision, where it ranks third behind MTRCNet-CL and SAHC. Specifically, for average accuracy, SPRMamba matches the high accuracy of LAST (a multi-task learning method that requires additional information in the form of extra labels) but displays a lower standard deviation of approximately 0.13%. Additionally, for average recall

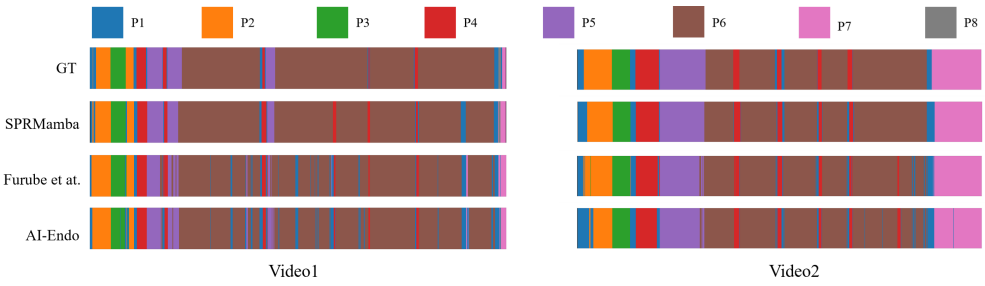

Figure 3: Qualitative comparisons with some other methods on ESD385 dataset. The following four rows show the ground-truth labels, the predictions by the proposed SPRMamba, the predictions by Furube et at., and the predictions by AI-Endo. P1 to P8 indicate the phase label.

and average Jaccard, SPRMamba achieves the best performance with 90.12% and 81.43%, respectively.

Furthermore, to illustrate the performance of our approach compared to state-of-the-art methods, in Fig. 4 we qualitatively compare two examples from the Cholec80 testing dataset. We compare the proposed method with two SOTA methods, TeCNO [7] and Trans-SVNet [9]. As shown in Fig. 4, our method predicts higher frame-level accuracy, especially on the boundary between two different phases. In addition, we can notice that Trans-SVNet and TeCNO make a lot of mistakes, *i.e.*, they are unable to accurately classify ambiguous frames in different phases. For example, some frames within P3 are misclassified into P4 and some frames within P5 are misclassified into P4 in Video 1. On the contrary, our proposed method performs very accurately compared to ground truth values.

4.5. Ablation Study

We conducted a series of ablation experiments on the ESD385 dataset to validate the effectiveness of each component and parameter setting in our proposed method on the model.

4.5.1. Scale Residual TranMamba

Table 4 shows the importance of the SRTM modules. Transformer and Mamba denote the use of transformer branches alone and Mamba branches alone, respectively. To keep the number of parameters constant, we still use $\mathbf{F} \in \mathbb{R}^{L\times 64}$ as input. SRTM improves the accuracy by 1.61% and 0.73%

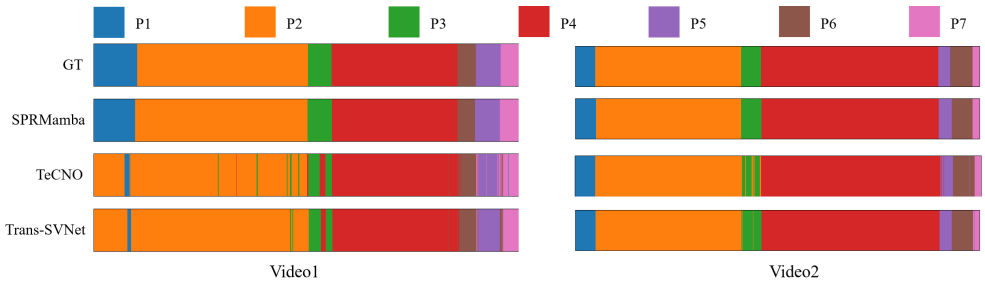

Figure 4: Qualitative comparisons with some other methods on Cholec80 dataset. The following four rows show the ground-truth labels, the predictions by the proposed SPRMamba, the predictions by TeCNO, and the predictions by Trans-SVNet. P1 to P7 indicate the phase label.

Table 4: ABLATION STUDY ON THE SCALE RESIDUAL TRANMamba. In the case of Transformer, we only use transformer branches instead of SRTM. In the case of Mamba, we only used the Mamba branch.

	Accuracy(%)	Precision(%)	Recall(%)	Jaccard(%)	FLOPs	Params
Transformer	86.03±10.89	86.02±17.40	84.89±17.03	75.44±18.41	8.5G	26.15M
Mamba	86.91±12.06	86.59±17.24	86.05±16.74	77.19±18.42	6.5G	25.28M
SRTM	**87.64±9.83**	**86.72±16.54**	**86.76±15.66**	**77.51±17.83**	7.5G	25.42M

compared to Transformer and Mamba, respectively, demonstrating the crucial role of modeling short-term and long-term temporal context for ESD surgical phase recognition.

4.5.2. Temporal Sample Strategy

Table 5: ABLATION STUDY ON THE TEMPORAL SAMPLE STRATEGY. In the case of STContext SRTM, we use two STContext SRTM blocks instead of a combination of STContext SRTM and LTContext SRTM. In the case of LTContext SRTM, we used two LTContext SRTM blocks.

	Accuracy(%)	Precision(%)	Recall(%)	Jaccard(%)
baseline	86.58±11.58	86.60±17.48	84.16±16.80	75.65±18.70
STContext SRTM	87.03±11.73	86.95±17.57	85.20±17.32	76.86±17.82
LTContext SRTM	87.38±11.15	**87.27±17.20**	85.42±16.43	77.07±17.89
Ours	**87.64±9.83**	86.72±16.54	**86.76±15.66**	**77.51±17.83**

To demonstrate the effectiveness of the Temporal Sample Strategy, we designed four ablation experiments. We also configured the original SRTM framework matching the SPRMamba structure as a benchmark for comparison. As shown in Table 5, baseline achieves an average accuracy of 86.58%, an average precision of 86.60%, an average recall of 84.16%, and an average

Jaccard of 75.65% for surgical phase recognition. The experimental results outperform the baseline modules regardless of whether the key modules are added individually or in combination with each other. First, we show the impact of combining STContext SRTM and LTContext SRTM in the LST-Context block shown in Fig 2. We compare it with two variants where only STContext SRTM or LTContext SRTM is used. To keep the number of parameters constant, in these cases, we still use both SRTM blocks in LST-Context. The results show that combining the two SRTM blocks gives better results.

4.5.3. Effect of Different value of W and G

Table 6: ABLATION STUDY ON DIFFERENT PARAMETERS W AND G.

W	G	Accuracy(%)	Precision(%)	Recall(%)	Jaccard(%)
8	64	87.53±11.10	**87.73±16.56**	85.39±16.21	77.41±17.16
16	64	87.37±9.63	86.56±16.23	86.23±16.30	77.35±17.17
32	64	86.91±11.52	87.62±17.82	85.10±15.87	76.80±18.23
64	64	**87.64±9.83**	86.72±16.54	**86.76±15.66**	**77.51±17.83**
64	32	87.33±10.41	86.50±16.74	85.62±16.32	76.85±17.65
64	16	87.08±10.98	87.09±17.19	85.15±16.42	76.77±17.91
64	8	87.11±10.76	86.35±18.76	85.88±17.40	76.60±18.76

To further explore the effect of the Temporal Sample Strategy with different W and G on the model performance, we conducted the experiments shown in Table 6. The parameter W controls the size of the local window and the parameter G controls the range of the long-term temporal context modeling. The experimental results show that the performance of the models (Accuracy and Jaccard) increases as G increases until the optimal performance is reached for G=64. The performance fluctuates as the size of the local window W becomes smaller, but W=64 works best.

4.5.4. Effect of using convolutions

Table 7: ABLATION STUDY ON USING CONVOLUTIONS

	Accuracy(%)	Precision(%)	Recall(%)	Jaccard(%)
Conv without Dilation	85.97±11.48	85.75±17.70	83.96±18.06	74.74±18.92
Without Conv	86.60±10.95	85.71±17.94	84.99±17.56	75.46±18.87
Conv with Dilation	**87.64±9.83**	**86.72±16.54**	**86.76±15.66**	**77.51±17.83**

To explore the impact of dilated convolution in LSTContext blocks on model performance. We compared 1D dilated convolution with 1D convolution using the same size but without dilation factor and without 1D convo-

lution. The results show that dilated convolution has a significant impact on performance.

5. Disscussion

The accurate recognition of surgical phases in ESD within CAS systems is crucial for enhancing surgical efficiency, minimizing patient complications, and providing valuable training material for novice endoscopists. However, considering the duration and complexity of the ESD process, the challenge of handling long sequences of video frames and effectively modeling temporal context with limited computing resources remains significant.

To address these challenges, we propose SPRMamba, a Mamba-based framework for online ESD surgical phase recognition. Our approach leverages the Scale Residual TranMamba module to effectively model both long-term and short-term temporal contexts, offering superior temporal modeling capabilities compared to traditional methods. Additionally, We introduce a novel Temporal Sample Strategy to mitigate the computational resource constraints, making the framework feasible for real-time surgical phase recognition. The advantages of our approach are as follows: 1) **Long-Term Modeling Superiority:** Mamba enhanced capabilities for long-term temporal modeling with lower computational complexity compared to Transformers. This makes it particularly suitable for handling the extended duration and complexity of ESD surgeries; 2) **Effectiveness of the SRTM Module:** The proposed SRTM module achieves the best performance due to its ability to capture complex phase relationships in ESD surgeries. While Mamba ensures overall phase recognition accuracy, the challenge of accurately recognizing short-duration phases remains, which our SRTM module addresses effectively; 3) **Advantages of the Temporal Sample Strategy** The proposed Temporal Sample Strategy outperforms direct long sequence processing and mitigates the secondary complexity typically associated with Transformer models, enhancing the efficiency of temporal context modeling and improving overall recognition performance. In addition, our proposed method allows for flexible modification of the temporal modeling length and specific adjustments for different surgical tasks.

Finally, our proposed method not only surpasses state-of-the-art (SOTA) methods in ESD surgical phase recognition but also demonstrates robustness for other surgical tasks. Specifically, our method achieves an average accuracy of 87.64%, an average precision of 86.72%, an average recall of

86.76%, and an average Jaccard of 77.51% on the ESD dataset, outperforming the next best method by significant margins. Additionally, validation on the cholecystectomy Cholec80 dataset further confirms our method's robustness, achieving superior results in most metrics compared to the SOTA methods[3, 45, 49, 26, 7, 6, 9, 48], except for precision. The qualitative results, as illustrated in Fig. 3 and Fig. 4, further substantiate the effectiveness and reliability of our approach.

Although our method shows high surgical phase recognition performance in ESD and cholecystectomy videos, it has some limitations. First, this study primarily focuses on ESD procedures, and while we validated the robustness of our method on the Cholec80 dataset, further experiments are needed to assess its generalizability across a broader range of surgical tasks. In the future, we plan to extend our method to other surgical video analysis tasks, potentially incorporating other innovations such as uncertainty analysis to enhance the robustness and accuracy of SPRMamba across different surgical. Second, the dataset used in this study was limited in size and diversity, which may affect the model's performance in more diverse clinical settings. Future work will focus on expanding the dataset to include multicenter data and a wider variety of surgical procedures.

6. Conclusion

In this paper, we address the critical need for accurate surgical phase recognition in Endoscopic Submucosal Dissection, a key procedure for treating early-stage gastrointestinal cancers. Achieving precise real-time phase recognition in ESD is essential for improving surgical outcomes and efficiency. However, traditional SPR methods face significant challenges, particularly in effectively capturing the temporal context over extended surgery durations and managing the high computational demands of video analysis required for real-time applications. To overcome these challenges, we propose SPRMamba, a novel framework designed to enhance the accuracy and efficiency of ESD surgical phase recognition. Specifically, our method proposes the Scale Residual TranMamba module, which combines the ability of Mamba to extract the long-term temporal context with the ability of the Transformer to extract the short-term temporal context and excels in capturing complex temporal relationships in surgical videos. In addition, considering the real-time requirements of surgical phase recognition, a temporal sampling strategy is designed to optimize computational resources by efficiently modeling

both short-term and long-term temporal contexts. Extensive experiments demonstrate that SPRMamba not only outperforms current state-of-the-art methods in accuracy and robustness but also significantly reduces the computational burden. In the future our approach has the potential to advance the field of surgical video analysis, offering a valuable tool for both clinical practice and surgical education.

Acknowledgment

This work was supported by the Joint Laboratory of Intelligent Digestive Endoscopy between Shanghai Jiaotong University and Shandong Weigao Hongrui Medical Technology Co., Ltd. and the National Science Foundation of China under Grant 62103263.

References

[1] T. R. McCarty, A. N. Bazarbashi, K. E. Hathorn, C. C. Thompson, H. Aihara, Endoscopic submucosal dissection (esd) versus transanal endoscopic microsurgery (tem) for treatment of rectal tumors: a comparative systematic review and meta-analysis, Surgical endoscopy 34 (2020) 1688–1695.

[2] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, et al., Surgical data science for next-generation interventions, Nature Biomedical Engineering 1 (9) (2017) 691–696.

[3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, N. Padoy, Endonet: a deep architecture for recognition tasks on laparoscopic videos, IEEE transactions on medical imaging 36 (1) (2016) 86–97.

[4] A. Huaulmé, P. Jannin, F. Reche, J.-L. Faucheron, A. Moreau-Gaudry, S. Voros, Offline identification of surgical deviations in laparoscopic rectopexy, Artificial Intelligence in Medicine 104 (2020) 101837.

[5] T. Vercauteren, M. Unberath, N. Padoy, N. Navab, Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions, Proceedings of the IEEE 108 (1) (2019) 198–214.