ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1266-1284
DOI 10.5755/j01.itc.55.2.44234

Endoscopic Surgical Procedure Recognition Based on Convolutional Networks and Feature Focusing

Received 2026/02/04 | Accepted after revision 2026/05/20

HOW TO CITE: Chen, L., Duan, F., Cui, P., Ling, L., Hu., J. (2026) Endoscopic Surgical Procedure Recognition Based on Convolutional Networks and Feature Focusing. *Information Technology and Control*, 55(3), 1266-1284. https://doi.org/10.5755/j01.itc.55.2.44234

Endoscopic Surgical Procedure Recognition Based on Convolutional Networks and Feature Focusing

Lei Chen*, Feiya Duan, Peigen Cui, Lin Ling, Jing Hu

School of Mechanical Engineering, Hefei University of Technology, Anhui, 230009, China; e-mails:
Lei Chen*: chenleiromeo@outlook.com, Feiya Duan: diff123@126.com, Peigen Cui: 15264750657@163.com
Lin Ling: linglin@hfut.edu.cn, Jing Hu: hj0551@hfut.edu.cn

Corresponding author: chenleiromeo@outlook.com

To address the issues of weak dynamism, feature redundancy, and noise interference in process recognition during endoscopic surgery, this paper proposes a deep learning framework integrating convolutional networks and feature focusing. This study designs a lightweight dual-channel feature extraction module, constructs a spatio-temporal attention-guided feature network, and combines dynamic threshold segmentation with multi-scale feature fusion. Experimental results show that, in terms of computational efficiency, the inference speed reaches 57fps, which is 40% higher than the traditional method; in regard to professional standardization, the average rule compliance rate remains at 89.3%. Under 50% image noise interference, the model accuracy remains at 88.7%, which can resist interference well in complex surgical scenarios. These data demonstrate the strong adaptability of the framework to complex surgical scenarios. This method can provide key technical support for intelligent surgical navigation and surgical quality assessment.

KEYWORDS: Endoscopic surgery; Surgical procedure recognition; Convolutional network; Feature focusing; Spatio-temporal attention

1. Introduction

Endoscopic surgery is an important surgical procedure. During the procedure, the instruments are complex to operate, the images are not dynamic, and lighting and occlusion can significantly interfere with the surgery [7, 21]. In recent times, convolutional networks (CNs) have been broadly applied for action recognition in complex scenes [1, 9]. At the same time, attention mechanisms and temporal modeling methods have also been adopted to assist in judging the surgical outcome [29]. However, these existing methods still have problems such as low recognition stability and insufficient noise robustness. To this end, this study proposes a Feature-Focused Convolutional Network for Surgical Phase Recognition (FocNet-SR) model. This model utilizes a lightweight dual-channel module to separate surgical instruments from tissue texture. It also combines an improved DeepLab Version 3 Plus (DeepLabV3+) network with a channel-spatial attention mechanism. The research aims to improve the precision of endoscopic surgical procedure recognition. The core innovation of this research lies in proposing a three-stage collaborative optimization paradigm, semantic decoupling, spatio-temporal focus, and decision smoothing, rather than incremental improvements to individual modules. It incorporates three novel elements distinct from prior approaches: 1) A dual-channel feature decoupling module. Unlike traditional methods that jointly model instrument and tissue textures, this module employs complementary attention masks to forcibly separate the two semantic entities, thereby suppressing feature contamination at source; 2) An enhanced Atrous Spatial Pyramid Pooling (ASPP) incorporating a Convolutional Block Attention Module (CBAM). Unlike the standard DeepLabV3+ approach which treats features at all scales equally, this enables the multi-scale extraction process to dynamically focus on critical surgical regions, achieving adaptive feature enhancement; 3) A dynamic threshold segmentation rule based on confidence sliding averages, which replaces fixed thresholds or simplistic post-processing. By dynamically adjusting segmentation boundaries through real-time calculation of historical confidence scores, it effectively filters out false phase transitions to ensure temporal consistency. These three components progressively synergise to systematically resolve three core challenges in complex surgical scenarios: feature contamination, ambiguous discrimination, and temporal jitter. This research aims to provide high-precision process recognition technology for intelligent surgery, and to offer reliable data support for clinical surgical quality assessment and decision support.

2. Literature Review

In recent years, image-based minimally invasive surgery has been widely adopted, and the key challenge in the surgical process has become the separation of instruments and human tissues. The accuracy of separation is crucial to the success of the surgery. To enable the automatic recognition of surgical procedures and the assessment of surgical skills during laparoscopic distal gastrectomy, Komatsu et al. [11] proposed a deep learning-based surgical phase recognition model. Trained using annotated surgical videos from multiple centers, this model accurately identified surgical stages and enabled the automatic assessment of surgical skills based on procedural features [11]. To address the issues of inefficient temporal modelling and the difficulty in capturing fine-grained phase transitions in the recognition of endoscopic submucosal dissection (ESD) phases, Zhang et al. [28] proposed the SPRMamba framework. By utilising the Mamba architecture and a hierarchical sampling strategy, it achieved efficient real-time surgical phase recognition, attaining state-of-the-art (SOTA) performance and demonstrating strong generalisation capabilities [28]. To address the performance bottlenecks inherent in the image pre-training and video fine-tuning paradigm, as well as the scarcity of surgical video data and the challenges of spatio-temporal modelling, Yang et al. [26] proposed the SurgPETL framework for efficient parameter transfer learning. Through an adaptive spatio-temporal representation modulation module, it achieved high-precision, lightweight, fine-grained surgical stage recognition [26]. To achieve precise surgical stage recognition for robot-assisted radical prostatectomy and validate its generalisation capa-

bility, Konnai et al. [12] proposed an AI recognition system based on the TeCNO network. Through annotation of surgical videos from multiple surgeons and validation via segmented modelling, the system achieved high-precision surgical stage recognition with cross-surgeon generalization performance [12].

Meanwhile, in the field of medical image analysis, CN-based methods have played a core role in improving the accuracy and robustness of lesion and tissue segmentation, especially performance optimization on multi-modal datasets has become a key focus of academic research.To improve the accuracy of retinal vessel segmentation, Muwei Jian et al. [10] proposed a dual-branch fully CN, U-Net.By enhancing feature extraction through parallel encoders, fusing convolutional down-sampling control information, and filtering noise through decoder attention modules, U-Net achieved segmentation performance superior to five existing methods on the DRIVE and ARIA datasets, promoting applications in medical image analysis and surgical procedure recognition [10]. To improve the accuracy and robustness of lesion segmentation in medical images, Liu et al. [14] proposed a cross-scale residual attention fully CN. By fusing multi-scale features through transformer-type cross-scale attention modules, enhancing representation through convolutional Kolmogorov-Arnold networks, and refining features through pointwise convolution, U-Net achieved significant improvements in F1 and cross-union ratio on the BUSI, PH2, and DDTI datasets [14]. To improve the comprehensive modeling ability of global and local features in medical image recognition, Dang et al. [5] proposed an efficient transformer of CN combined with edge constraints. By extracting features through parallel CN and Transformer branch, and using the saliency module to supervise edge information, a significant improvement in Dice score was achieved [5]. To improve the segmentation accuracy of chest X-ray and dermoscopic medical images, Nagendram et al. [18] proposed an ensemble approach grounded in deep CN. By combining U-Net and fully CN with stochastic gradient descent optimization, high sensitivity, accuracy and Dice coefficient segmentation performance were achieved on PH2 and chest X-ray datasets [18].

While the aforementioned studies have made progress in static image segmentation of specific tissues, they have failed to effectively address the core challenges of dynamic process recognition in endoscopic surgery. These methods generally lack explicit modeling of the semantic interaction between the instrument and tissue, leading to feature mixing and insufficient discriminative power in complex surgical scenarios. They also neglect long-term temporal dependencies and clinical rule constraints, have limited robustness to intraoperative dynamic disturbances, and their computational efficiency is insufficient for real-time requirements. Therefore, this study proposes a recognition framework that integrates CNs and feature focusing. A lightweight dual-channel feature decoupling module is designed to achieve semantic separation between instruments and tissues; a spatio-temporal attention network was constructed to enhance the discrimination and temporal coherence of key steps. A dynamic decision optimization strategy was combined to improve the overall performance and clinical applicability of the system in complex dynamic scenarios.

3. Methodology

3.1. Endoscopic Image Preprocessing and Multi-scale Feature Extraction

The quality of endoscopic surgical video images is limited by complex lighting conditions, tissue reflection characteristics and intraoperative interference (such as blood and smoke) inside the body. Directly processing the original images will seriously affect the performance and stability of subsequent DL models [27]. Therefore, it is crucial to construct a robust feature extraction front-end. To improve the accuracy of subsequent feature extraction, the study first performs systematic preprocessing on the images to provide high-quality input for the CN. The endoscopic image preprocessing process is shown in Figure 1.

As shown in Figure 1, the original endoscopic image is first subjected to gamma correction to expand the dynamic range of the dark area to improve the overall contrast. Then, contrast-limited adaptive histogram equalization is applied to divide the image into 8×8 local regions and limit the contrast enhancement, thereby enhancing the local texture of the tissue while suppressing noise amplification. Subsequent-

Figure 1
Schematic diagram of endoscopic image preprocessing.

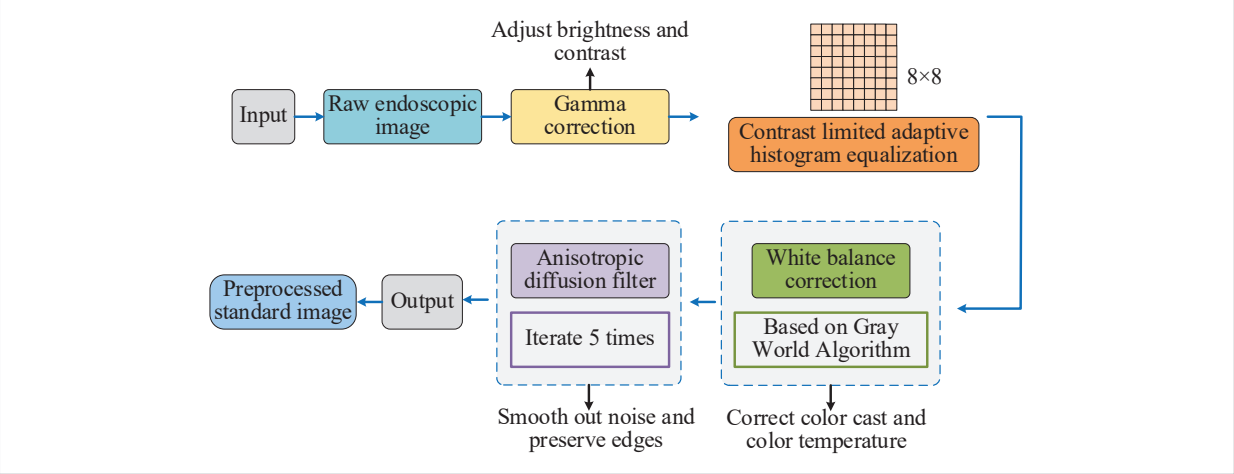
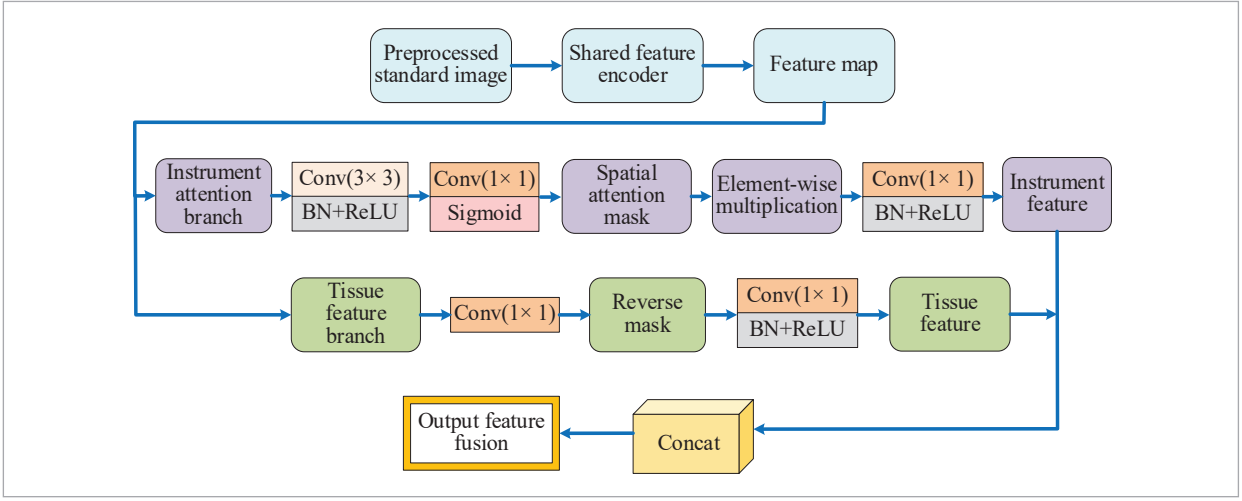


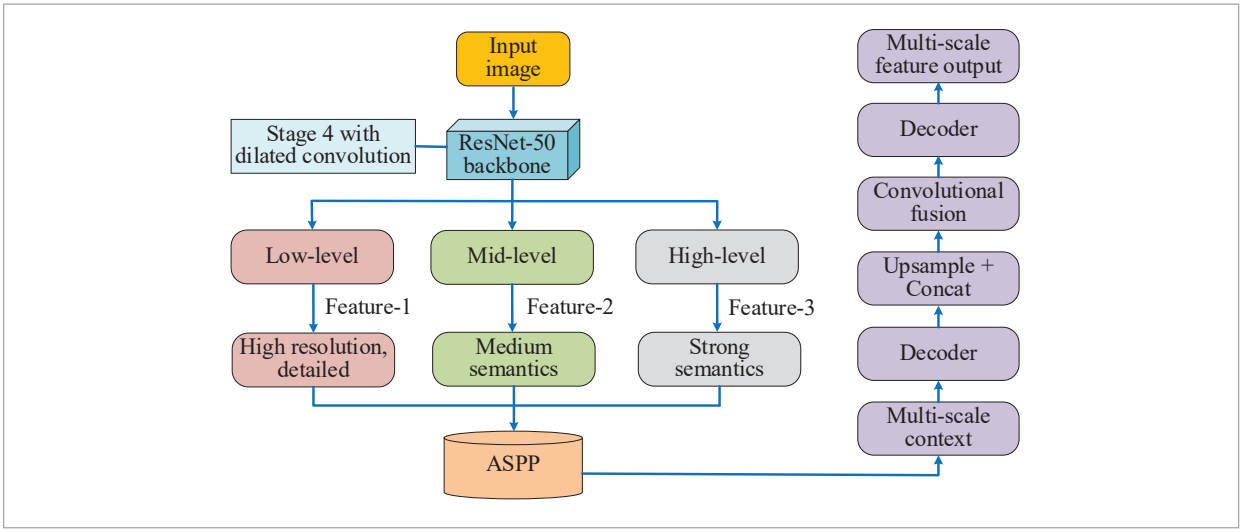
Figure 2
Schematic diagram of dual-channel feature decoupling module.



ly, a white balance algorithm based on the gray world hypothesis is used for color correction to eliminate the color shift caused by the characteristics of the endoscopic light source or the reflection of tissue fluid and restore the natural color of the tissue [17]. Finally, anisotropic diffusion filtering is applied for 5 iterations to smooth the noise in the homogeneous region while protecting and enhancing important edge information. This series of sequential processing operations together transform the original endoscopic image into a standardized, high-quality feature extraction input, providing a stable and reliable visual foundation for subsequent DL models. However, endoscopic images contain two significantly different semantic entities: surgical instruments and human tissue. If mixed processing is performed, their features are prone to mutual interference. Therefore, a lightweight feature decoupling module is designed, as shown in Figure 2.

As shown in Figure 2, the module first uses a lightweight shared encoder to extract basic features. Then, the network is divided into two parallel branches. The surgical instrument attention branch generates a spatial attention mask through a small

Figure 3
Schematic diagram of DeepLabV3+ multi-scale feature extraction architecture.



CN to identify the probability that each position in the image belongs to a surgical instrument [6]. The generation process of mask M_{tool} can be formalized as shown in Equation (1).

$$M_{tool} = \sigma(F_{mask}(F_{base};\theta_{mask})) . \quad (1)$$

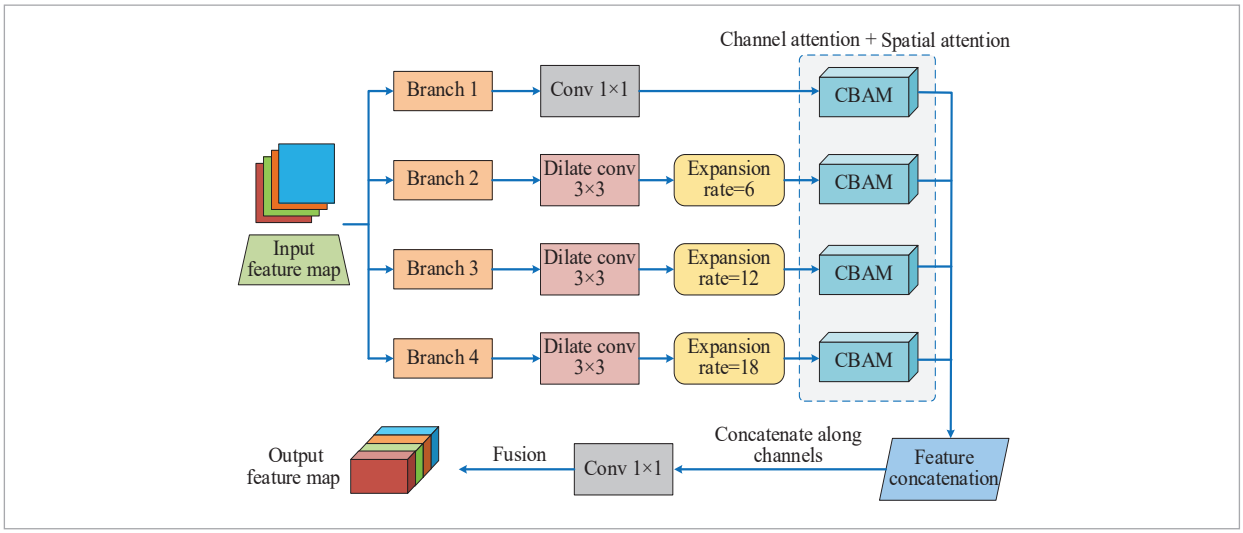
In Equation (1), F_{base} represents the basic feature; $Fmask(\cdot;\theta mask)$ represents the lightweight convolutional sub-network with parameter θ_{mask}; σ represents the Sigmoid activation function, which normalizes the output to the 0~1 range. By multiplying this mask with F_{base}, spatial filtering and enhancement of instrument features are achieved. The tissue feature branch adopts a complementary strategy, using an inverse mask to weight F_{base}, forcing the network to focus on extracting tissue texture and anatomical structure information. The feature decoupling and extraction process of the two branches can be uniformly represented as shown in Equation (2).

$$\begin{cases} F_{tool} = F_{tool}(F_{base} \odot M_{tool}) \\ F_{tissue} = F_{tissue}(F_{base} \odot (1 - M_{tool})) \end{cases} . \quad (2)$$

In Equation (2), F_{tool} and F_{tissue} represent the output features of the tool and tissue branches, respectively; $\odot$ represents element-wise multiplication; $Ftool(\cdot)$ and $Ftissue(\cdot)$ represent the 1×1 convolutional layers used to refine the tool and tissue features, respectively. Finally, the output features of the two branches are concatenated in the channel dimension to form semantically separated but interconnected coupled features. After obtaining the semantically decoupled coupled features, it is necessary to further capture rich spatial context information from them to cope with the drastic changes in target scale in the surgical scene. To this end, the study introduced a multi-scale feature extraction structure based on DeepLabV3+ CN, as shown in Figure 3 [20].

As presented in Figure 3, the study uses ResNet-50 with dilated convolution as the backbone network, which obtains a wide receptive field while maintaining a high spatial resolution of the feature map [16]. The backbone network naturally outputs multi-level features: low-level features have high resolution and are rich in edge and texture details; high-level features have strong semantic information but sparse spatial details. Then, the ASPP module is applied to the high-level features, and multi-scale context features are captured through parallel multi-rate dilated convolution. Subsequently, multi-scale feature fusion is performed through a decoder path: the multi-scale context features output after ASPP processing are upsampled and then concatenated and fused with the low-level detail features from the early stage of the backbone network and processed, and finally multi-scale fused features are output [30]. The

Figure 4
Schematic diagram of the improved ASPP module.



effective aggregation of multi-scale context information depends on the design of the ASPP module. The standard ASPP treats all scale information equally and fails to consider the differences in their importance to the current recognition task [23]. To better adapt to the dynamic changes in target scale and importance in surgical scenarios, key enhancements are made to the ASPP module, as shown in Figure 4.

As presented in Figure 4, the core innovation of the ASPP module lies in the introduction of the CBAM. A CBAM attention module is inserted after each of the three dilated convolution branches with different dilation rates. CBAM is a lightweight general attention module that calculates attention maps in an orderly manner along the channel and spatial dimensions [2]. Channel attention first performs global average pooling (GAP) and max pooling (MP) on the input features, and then generates channel weights through a shared multilayer perceptron [24]. Given intermediate feature $X \in R^{H\times W\times C}$, the calculation of its channel attention map $M_c \in R^{1\times 1\times C}$ is shown in Equation (3).

$$M_c = \sigma(W_1(W_0(\mathbf{z}_{avg})) + W_1(W_0(\mathbf{z}_{max}))) \quad (3)$$

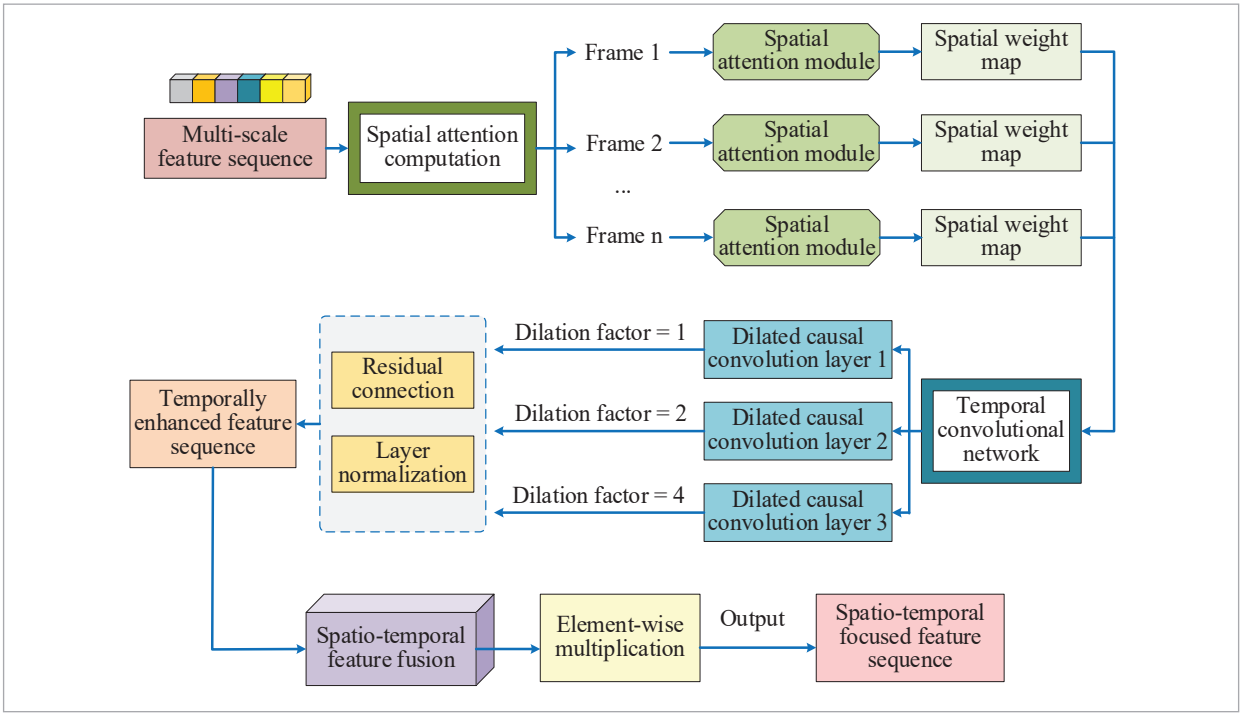
In Equation (3), $\mathbf{z}_{avg}$ and $\mathbf{z}_{max}$ represent the GAP feature and the MP feature, respectively; W_0 and W represent the weights of the two fully connected layers of the MLP, respectively. Spatial attention generates a spatial weight map by merging the GAP and MP information of the channel dimension, which is used to highlight important spatial regions in the feature map [19]. The calculation of the spatial attention map M_s is shown in Equation (4).

$$M_s = \sigma(f^{7\times 7}([\text{AvgPool}(X); \text{MaxPool}(X)])). \quad (4)$$

In Equation (4), X represents the feature tensor with dimension $H\times W\times C$ (where H, W, and C represent the height, width, and number of channels of the feature map, respectively); $[\cdot]$ represents the concatenation along the channel dimension; AvgPool(X) represents the global GAP operation along the channel dimension; MaxPool(X) represents the global MP operation along the channel dimension; and $f^{7\times 7}$ represents a 7×7 convolution operation, used to generate spatial attention weights. The integration of channel and spatial attention in CBAM is particularly well-suited for distinguishing instruments from tissue in endoscopic images. Channel attention captures the global response across different feature channels: surgical instruments typically exhibit concentrated color distributions due to metallic materials (e.g., the silvery-white hue of forceps), corresponding to specific channel activation patterns. Spatial attention, meanwhile, focuses on localized regions: complex and diverse tissue textures (e.g., blood vessels,

Figure 5

A feature focusing network structure diagram guided by spatio-temporal attention.



adipose tissue) often concentrate discriminative information at specific spatial locations (e.g., the operative site). The synergistic interaction between channel attention and spatial attention works as follows: channel attention filters the feature dimensions related to the color of the instruments, while spatial attention locates the key areas of tissue textures. This synergy enables the network to precisely differentiate the semantic information of surgical instruments and tissues along two dimensions: "what color/material it is" and "where it is located". Finally, the outputs of all branches are concatenated along the channel dimension and fused and dimensionality reduced by a 1×1 convolution to generate the final multi-scale contextual features, which significantly improves feature discriminativeness and the ability to focus on key information.

3.2. Optimization of Dynamic Surgical Procedure Recognition Model

After image preprocessing and multi-scale feature extraction, robust and focused multi-scale spatial features of each frame were obtained. However, the essence of surgical procedure recognition is to understand a dynamic, temporally logical sequence of events. Therefore, this study utilizes these high-quality frame features to construct an optimized model capable of modeling spatio-temporal dependencies and outputting a smooth, clinically compliant sequence of stages. The overall structure of the spatio-temporally attention-guided feature-focusing network constructed in this study is shown in Figure 5.

As presented in Figure 5, the network consists of two parts: parallel spatial attention and serial temporal modeling. First, the spatial attention module acts independently on the features of each frame. This module generates a two-dimensional spatial weight map W_s^t through a lightweight convolutional sub-network, whose value represents the importance of each position on the feature map for the current frame flow recognition [8]. The calculation of the weight map W_s^t can be expressed as shown in Equation (5).

$$W_s^t = \text{Sigmoid}(F_{spatial}(F_{multi}^t; \theta_{spatial})) . \quad (5)$$

In Equation (5), $F_{spatial}(\cdot;\theta_{spatial})$ represents the spatial attention sub-network with parameter $\theta_{spatial}$; F_{multi}^t

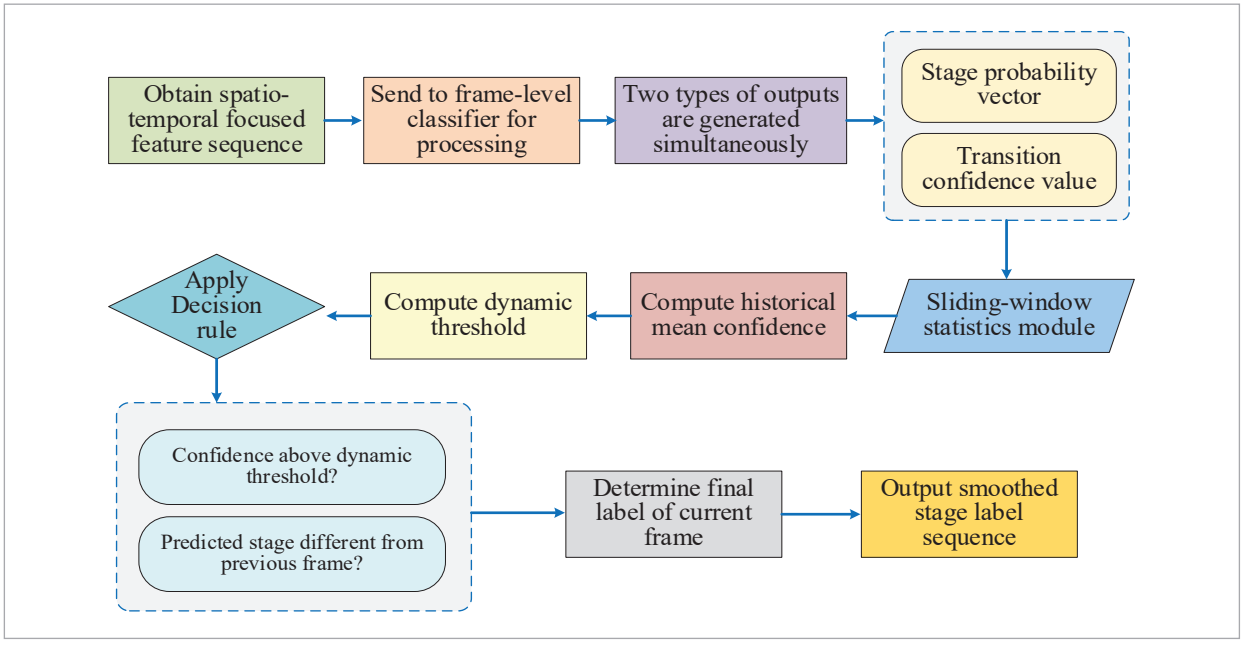
represents the feature of each frame. This simulates the process by which a surgeon focuses their visual attention on key areas of the surgical field (such as the operation point) at a specific moment. The spatial weight maps of all frames, together with the original features, are fed into a Temporal Convolutional Network (TCN) for modeling. The TCN is composed of multiple stacked residual blocks, each containing dilated causal convolutions. The dilation factor increases exponentially with the number of layers, thereby ensuring that the prediction of the current frame depends only on past frames and has a sufficiently large temporal receptive field [13, 25]. The operation of the one-dimensional dilated causal convolution in the TCN at the temporal level t can be formalized as shown in Equation (6) [3].

$$\mathbf{h}_t = \sum_{k=0}^{K-1} \mathbf{w}_k \cdot \mathbf{x}_{t-d\cdot k} \,. \tag{6}$$

In Equation (6), $\mathbf{h}_t$ represents the output vector at time step t; k represents the position index in convolution kernel 1; $\mathbf{x}$ represents the input sequence; $\mathbf{w}_k$ represents the weight vector of the convolution kernel at the position k; d represents the dilation factor; K represents the size of the convolution kernel; $\mathbf{x}_{t-d\cdot k}$ represents the value of the input sequence at time step $t–d\cdot k$. The design of exponentially increasing expansion factors with layer depth aligns with the multiscale temporal structure of surgical phases: shallow layers with small factors (d=1,2) maintain high temporal resolution to precisely capture rapid phase transitions like “clamping”; deep layers with large factors (d=4,8) extend the receptive field to minute-level scales, modeling long-duration phase dependencies such as “dissection.” This exponential scaling enables the network to cover cross-scale temporal windows from seconds to minutes with logarithmic depth, balancing computational efficiency with the distribution characteristics of phase durations. By stacking multiple such layers, TCN can capture long-range temporal dependencies. Finally, a spatio-temporal focusing operation is performed: the temporal enhancement features of each frame are multiplied element-wise with their corresponding spatial attention weight map [15]. Based on the spatio-temporal features after focusing, the model needs to perform the final stage classification. To avoid the result jitter caused by single-frame classification, a post-processing optimization strategy, namely dynamic threshold segmentation, is designed, as shown in Figure 6.

Figure 6

Dynamic threshold segmentation diagram.



As shown in Figure 6, the strategy begins with a spatio-temporal focused feature sequence. First, each frame feature is processed by a fully connected classifier, which outputs two key pieces of information: stage probability distribution (indicating the probability that the current frame belongs to each possible surgical stage) and transition confidence (a scalar between 0 and 1, reflecting the model's confidence in the current prediction). Subsequently, the system maintains a sliding time window, calculates the historical mean of the confidence of each frame within the window in real time, and calculates a dynamic threshold based on this mean. This threshold will be adaptively adjusted as the prediction stability changes [4]. The formula for calculating the dynamic threshold T_t is shown in Equation (7) [22].

$$T_t = \alpha \cdot \mu_t + \beta . \tag{7}$$

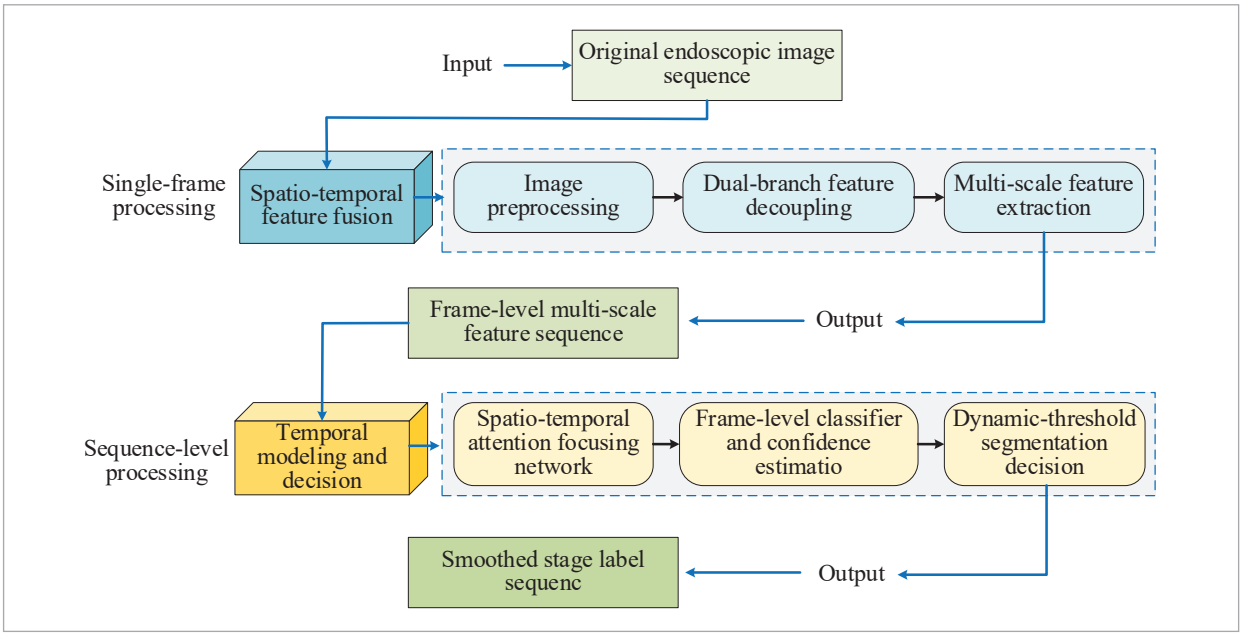
In Equation (7), α and β represent hyperparameters, which are used to control the sensitivity of the threshold and the baseline, respectively; μ_t represents the moving average of the confidence level calculated at the time step t, and its calculation is shown in Equation (8).

$$\mu_t = \frac{1}{|W_t|} \sum_{c_i \in W_t} c_i . \tag{8}$$

In Equation (8), W_t represents the sliding window at time step t; c_i represents the i-th confidence value within the window. Finally, the decision module applies a dual-condition judgment: only when the transition confidence of the current frame exceeds the dynamic threshold and the stage category predicted by the classifier is different from the output label of the previous frame, is a true stage transition determined and the current frame label updated; otherwise, the current frame label will remain consistent with the previous frame. Unlike traditional methods with fixed thresholds or moving averages that suffer from blurred boundaries, dynamic threshold segmentation offers adaptability and boundary preservation: it dynamically adjusts decision criteria by calculating the moving average of confidence scores in real time, raising thresholds during fluctuations to suppress false positives and lowering them during stable periods to maintain sensitivity. Combining dual thresholds for "confidence exceeding the threshold" and "category change," it smooths the time series while precisely preserving true phase

Figure 7

Overall framework diagram of FocNet-SR endoscopic surgical procedure recognition.



transition boundaries. This approach is particularly suited for surgical recognition scenarios characterized by significant phase duration variations and ambiguous transition boundaries. Finally, all the above core technology modules are integrated to construct a complete framework for the Feature-Focused Convolutional Network for Surgical Phase Recognition (FocNet-SR) model based on CNs and feature focusing, as shown in Figure 7.

As shown in Figure 7, the entire system takes the original endoscopic video stream as input and first acquires an image sequence through uniform frame sampling. The processing flow is then clearly divided into two main stages: spatial feature extraction and temporal modeling decision-making. In the first stage, each frame of image sequentially undergoes three core modules: preprocessing, dual-channel feature decoupling, and multi-scale feature extraction based on an improved DeepLabV3+. This aims to separate and enhance the semantic information of instruments and tissues from complex single-frame images, ultimately outputting a discriminative frame-level feature sequence. In the second stage, this feature sequence is first focused by a spatio-temporal attention network, fusing spatial key region information and temporal contextual dependencies. Then, a frame-level classifier generates preliminary probability predictions and confidence scores. Finally, a dynamic threshold segmentation decision-making module smooths the sequence, outputting a stable and coherent surgical procedure stage label sequence. Each part of this architecture plays its own role and provides support for the next part, presenting a recursive relationship that ensures the dynamic progress of the surgery and improves the precision and adjustability of the surgery.

4. Results and Analysis

4.1. Endoscopic Surgical Procedure Recognition Performance Test

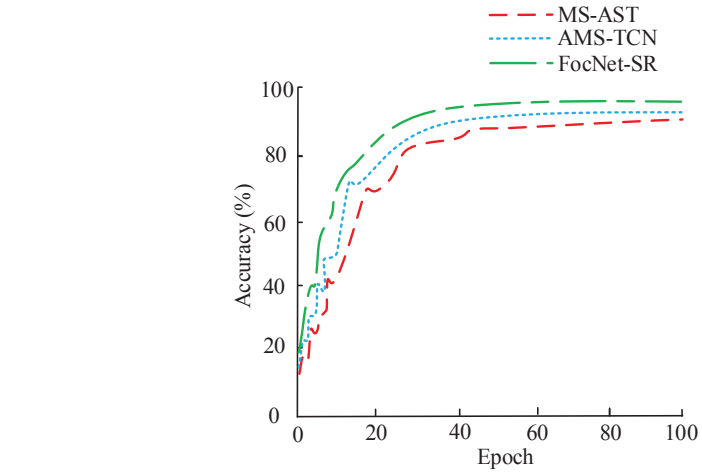
The experiments were conducted on a high-performance DL platform to ensure the training and inference efficiency of the endoscopic surgery procedure recognition model. Model training used the Adam optimizer with an initial learning rate of 0.0001, a batch size of 8, and a total of 100 epochs, employing cosine annealing learning rate scheduling and early stopping strategies. Data augmentation included random flipping, brightness and contrast adjustment, and rotation. The dynamic threshold segmentation hyperparameters α and β were set to 1.5 and 0.1, and the sliding window size was set to 7 frames. The Cholec80 dataset contained 80 videos of laparoscopic cholecystectomy. Each frame was finely annotated with seven standard surgical stages: preparation, exposure, clamping, transection, dissection, hemostasis, and removal. To eliminate sequential bias and ensure comparability, the study adopted a dual-track validation strategy: 1) Standard fixed partitioning (videos 41–80 as the test set), the most widely-used benchmark in the Cholec80 dataset, ensuring direct comparability with existing literature; 2) 4-fold cross-validation, randomly dividing the 80 videos equally and reporting the mean ± standard deviation across four experiments to quantify model stability. Both approaches strictly isolated training and test sets with no data leakage, balancing comparability and reliability of results. The specific experimental setup is shown in Table 1.

To verify the superiority of the proposed FocNet-SR model, this study introduced two mainstream models for comparison: the Multi-Scale Action Seg-

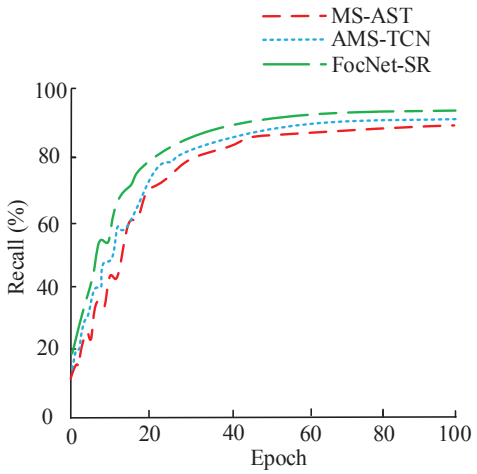
Table 1
Detailed experimental environment settings.

Type	Name	Model/Specification	Type	Name	Model/Specification
Hardware	GPU	NVIDIA RTX 4090	Software	Operating System	Ubuntu 22.04
Hardware	CPUq	Intel i9-13900K	Software	DL Framework	PyTorch 2.1
Hardware	Memory	128 GB DDR5	Software	GPU Acceleration	CUDA 12.2, cuDNN 8.8
Hardware	Storage	2 TB NVMe SSD	Software	Optimizer	AdamW
/	/	/	Software	Loss Function	Weighted Cross-Entropy

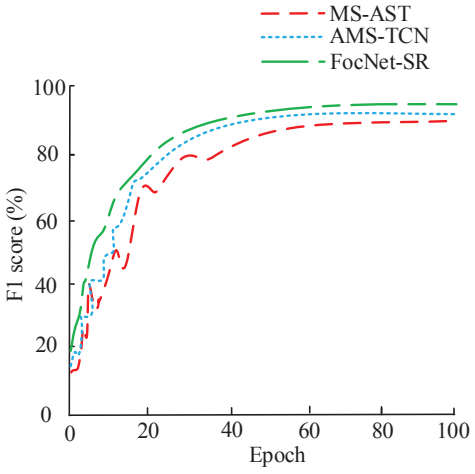
Figure 8
Comparison of accuracy, recall, and F1 score among the three models under various training cycles.



(a) Comparison of accuracy of four models under different training cycles



(b) Comparison of recall of four models under different training cycles

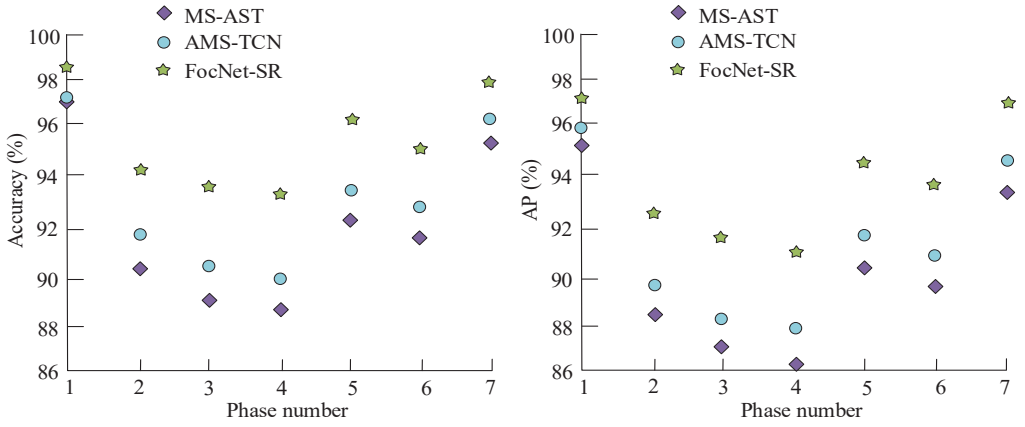


(c) Comparison of F1 score of four models under different training cycles

mentation Transformer (MS-AST) and the Adaptive Multi-Scale Temporal Convolutional Network (AMS-TCN). Model performance was evaluated using three metrics: accuracy (the proportion of correctly predicted frames), recall (the proportion of positive examples correctly predicted), and F1 score (the harmonic mean of accuracy and recall). The results are shown in Figure 8.

As presented in Figure 8, FocNet-SR achieved the best performance across all three evaluation metrics. This indicated that it had the best ability to perform timeline-based analysis and exhibited more stable convergence. As shown in Figure 8(a), within 20 rounds, the accuracy of FocNet-SR quickly reached over 70%, while MS-AST and AMS-TCN only achieved 60%–65%. FocNet-SR's final accuracy reached 95.8% after 100 rounds, surpassing the other two models. In Figure 8(b), FocNet-SR's recall exceeded 94% after 40 rounds, eventually stabilizing at 94.7%. The spatio-temporal attention guidance mechanism employed by FocNet-SR improves its performance in capturing dynamic images. In Fig-

Figure 9
Comparison of accuracy and AP among three models across seven surgical stages.



(a) Comparison of accuracy of three models in seven surgical stages

(b) Comparison of AP of three models in seven surgical stages

Table 2
Per-phase precision, recall, and F1 score of different models.

Phase	Metric	MS-AST	AMS-TCN	FocNet-SR
Preparation	Precision/Recall/F1	92.3/90.1/91.2	93.5/91.8/92.6	96.2/95.4/95.8
Exposure	Precision/Recall/F1	89.5/87.2/88.3	91.0/89.5/90.2	94.7/93.9/94.3
Clamping	Precision/Recall/F1	86.1/84.5/85.3	88.7/87.1/87.9	93.2/92.5/92.8
Transection	Precision/Recall/F1	85.3/83.8/84.5	87.9/86.2/87.0	92.8/91.9/92.3
Dissection	Precision/Recall/F1	90.2/88.9/89.5	92.1/90.8/91.4	95.3/94.7/95.0
Hemostasis	Precision/Recall/F1	84.7/82.9/83.8	87.3/85.6/86.4	91.5/90.8/91.1
Removal	Precision/Recall/F1	91.8/90.5/91.1	93.2/91.9/92.5	96.1/95.3/95.7

ure 8(c), FocNet-SR achieved an F1 score of 95.2% after 100 rounds, which was attributed to the threshold segmentation strategy in the model.

The study then tested the accuracy and average precision (AP) of each model in seven standard surgical stages. AP denotes the area under the precision-recall curve, comprehensively reflecting the model's recognition stability across different stages. The results are shown in Figure 9.

In Figure 9(a), FocNet-SR achieved the highest performance across all seven surgical stages, with an average stage accuracy of 95.6%. This was directly attributable to its dual-channel feature decoupling design. In Figure 9(b), FocNet-SR's average AP reached 94.0%, maintaining a high AP value even in complex stages. This indicated that its spatio-temporal attention network accurately characterized the confidence evolution throughout the entire action process. By combining feature focusing with spatio-temporal modeling, FocNet-SR achieved accurate and stable recognition of complex surgical procedures. To further evaluate the models' fine-grained recognition performance across surgical stages, the precision, recall, and F1 scores of FocNet-SR and the comparison models were calculated for seven stages, with results presented in Table 2.

As shown in Table 2, FocNet-SR outperformed the comparison models across all three metrics at every stage. Particularly in the most challenging occlusion and disconnection phases, its F1 scores reached 92.8% and 92.3% respectively, representing an approximately 5% improvement over AMS-TCN. This demonstrated that its dual-channel decoupling and spatio-temporal focus design effectively enhanced recognition robustness in complex interactive scenarios. This study further compared the computational efficiency and resource consumption of the traditional baseline method, Two-Stream Convolutional Neural Networks (TS-CNN), under unified hardware platform and standard testing conditions. Key performance indicators are presented in Table 3.

In terms of model complexity, FocNet-SR had 31.2 M parameters and 23.5 GFLOPs of computation. Regarding inference speed, the proposed FocNet-SR model achieved 57.0 fps. The single-frame processing latency of FocNet-SR was only 17.5 ms. The Transformer-based MS-AST model had the longest inference latency, reaching 28.4 ms. In summary, FocNet-SR achieved an excellent balance between model accuracy, inference speed, and resource overhead.

Finally, a systematic ablation experiment was conducted to verify the contributions of each core

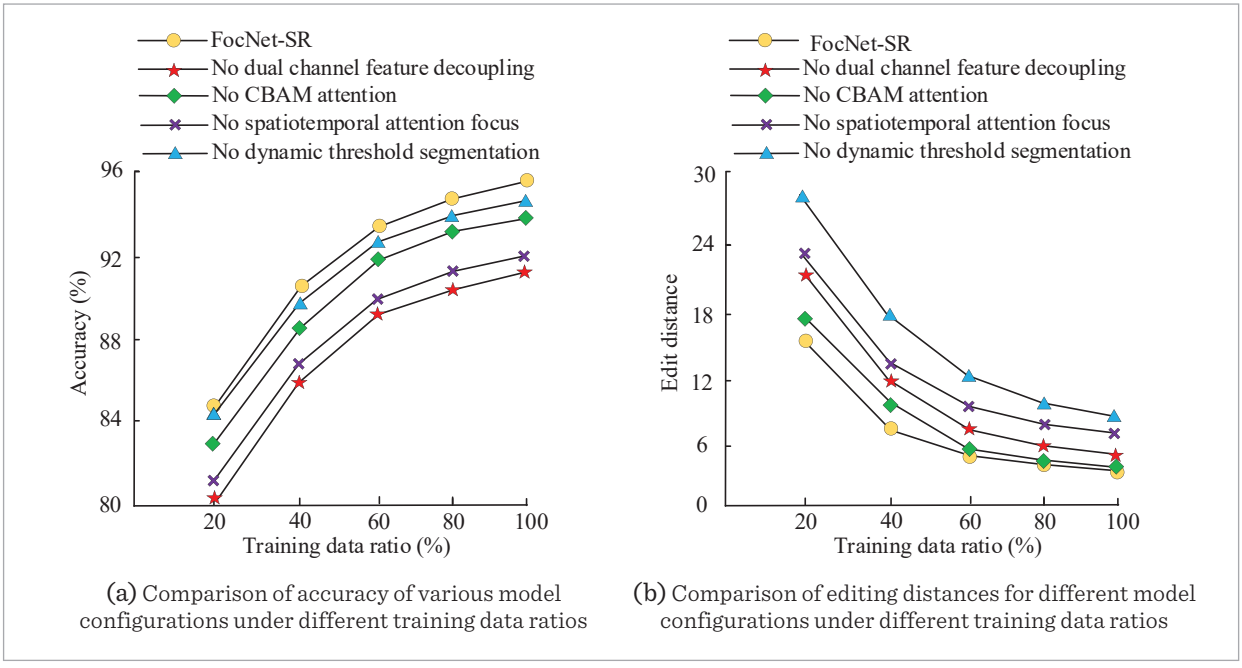
Table 3

Comparison of computational efficiency and resource consumption.

Method	Parameters (M)	Computation (GFLOPs)	Inference speed (fps)	Inference latency (ms)	Memory usage (GB)	Training time (hours)
TS-CNN	45.2	32.8	4 0.7	24.4	4.5	12.8
MS-AST	62.5	41.6	35.2	28.4	6.2	18.5
AMS-TCN	29.4	22.1	50.3	19.9	3.1	14.2
FocNet-SR	31.2	23.5	57.0	17.5	3.3	16.7

Figure 10

Ablation experiment comparison.



(a) Comparison of accuracy of various model configurations under different training data ratios

(b) Comparison of editing distances for different model configurations under different training data ratios

module of FocNet-SR. Five model variants were designed for comparison: 1) the complete FocNet-SR model; 2) the model with the dual-channel feature decoupling module removed; 3) the model with the CBAM attention mechanism removed from ASPP; 4) the model with the spatio-temporal attention-guided feature focusing network removed; and 5) the model with the dynamic threshold segmentation decision layer removed. Frame-level recognition accuracy was assessed using precision, while edit distance measured the degree of discrepancy between the predicted sequence and the actual sequence during the prediction phase. A lower value indicates superior temporal coherence. The results are shown in Figure 10.

As shown in Figure 10(a), with 40% training data, the accuracy of the complete model reached 90.6%, which was about 2% higher than the model without CBAM attention. This indicated that channel and spatial attention were particularly important for identifying key surgical steps in multi-scale feature focusing. The accuracy of the model without dual-channel feature decoupling was 86.2%, slightly lower than other ablation versions, indicating that feature decoupling helped to separate and distinguish instrument and tissue information. Figure 10(b) shows that the edit distance decreased with the increase of training data. The edit distance without dynamic threshold segmentation was consistently high, especially in the low data volume stage, indicating that dynamic thresholding has a significant effect on smoothing process identification and filtering out transient noise.

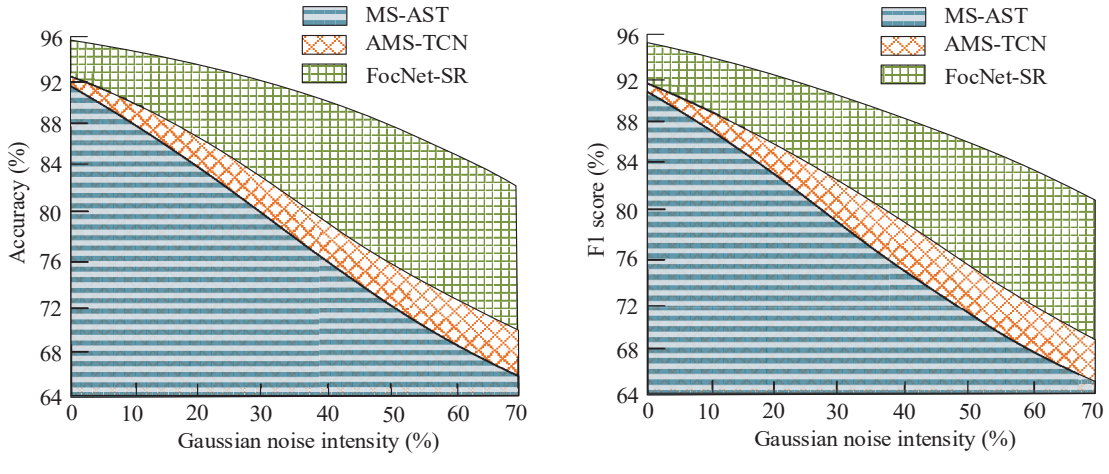
4.2. Simulation Verification of Endoscopic Surgical Procedure Recognition

To evaluate the robustness of the FocNet-SR model in the complex environment of real surgery, various simulation experiments were designed to simulate different interference situations that may be encountered in actual surgery. First, Gaussian noise of different intensities was added to the test set images to simulate image quality degradation caused by factors such as motion and lighting changes during surgery. The noise intensity was expressed as a percentage of the noise variance to the maximum gray value of the image. The experimental outcomes are presented in Figure 11.

As presented in Figure 11(a), as the Gaussian noise intensity gradually increased from 0% to 70%, the recognition accuracy of the three models generally showed a continuous downward trend, but FocNet-SR maintained a significant advantage at all noise levels. For example, at 20% noise, its accuracy was still 91.1%, which was about 7.6% and 5.9% higher

Figure 11

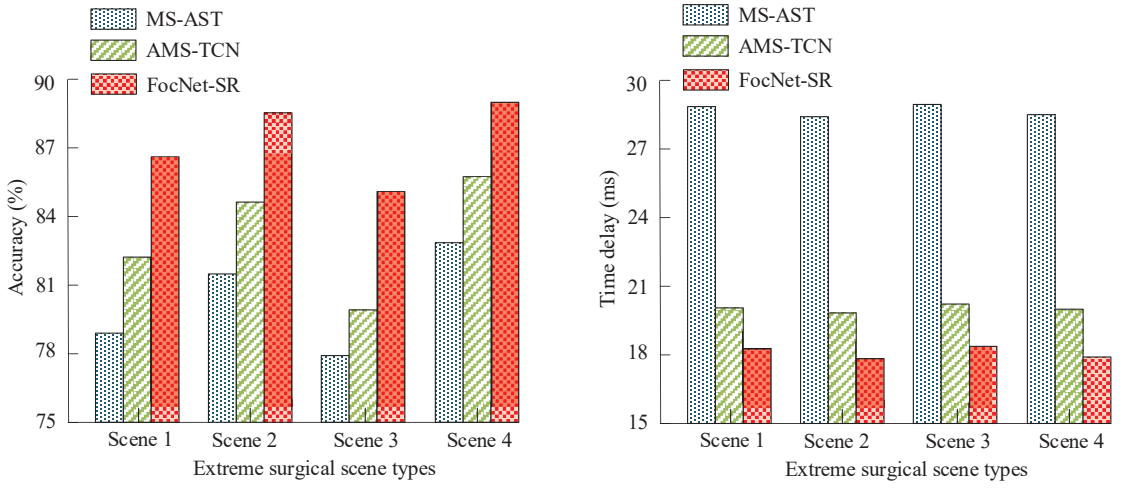
Comparison of recognition accuracy and F1 score of various methods under different Gaussian noise intensity.



(a) Comparison of recognition accuracy of various models under different Gaussian noise intensities

(b) Comparison of recognition F1 score of various models under different Gaussian noise intensities

Figure 12
Robustness comparison in extreme surgical scenarios.



(a) Comparison of recognition accuracy of three models in extreme surgical scenarios

(b) Comparison of time delay of three models in extreme surgical scenarios

than MS-AST and AMS-TCN, respectively, indicating that its feature robustness and noise resistance were stronger. At 50% noise intensity, its accuracy remained at 88.7%, significantly higher than AMS-TCN's 76.3% and MS-AST's 72.8%. This was mainly due to its dual-channel feature decoupling module, which could effectively separate the semantic information of instruments and tissues under noise interference, maintaining the discriminative power of basic features. In Figure 11(b), when the noise intensity increased from 0% to 70%, the F1 score of FocNet-SR decreased by only 13.1%, a much smaller decrease than that of the comparison model. This was attributed to its spatio-temporal attention network being able to focus on key spatio-temporal features in dynamic images, while the dynamic thresholding strategy ensured the temporal consistency of the output sequence, thus balancing precision and recall. As shown in Figure 11, FocNet-SR achieved end-to-end noise robustness improvement from feature extraction to final decision through the collaborative design of feature decoupling, attention focusing, and decision optimization, providing a solid guarantee for model deployment in real-world surgical scenarios. This anti-interference design philosophy aligned closely with the systematic disturbance defence framework proposed in [7]. This framework posits that a robust visual system should comprise three tiers: disturbance modelling, feature robustness enhancement, and decision correction mechanisms. The front-end preprocessing and dual-channel decoupling modules correspond to 'perturbation modelling + feature enhancement' – simulating noise distributions through image enhancement and suppressing semantic confusion via feature separation. The dynamic threshold segmentation strategy corresponds to 'decision correction' – filtering noise-induced pseudo-transitions through adaptive thresholds in the temporal output layer. Together, these form a complete closed-loop from 'feature enhancement' to 'decision defence', elevating the research methodology from isolated feature-level robustness to a system-level anti-interference framework. Based on this, the study tested the model's recognition performance under four extreme surgical scenarios: significant bleeding occlusion (Scenario 1), rapid instrument movement blur (Scenario 2), smoke interference (Scenario 3), and tissue fluid reflection (Scenario 4), as shown in Figure 12.

Figure 12(a), FocNet-SR exhibited the best robustness under all four challenging conditions, with an average accuracy of 87.4%. In the "rapid movement of equipment blur" scenario, its accuracy reached

88.5%, which was about 7.1% and 3.9% higher than MS-AST and AMS-TCN, respectively, indicating that the model could still stably extract structural features under motion blur and instantaneous information loss. This advantage aligned with the dynamic blur processing and feature retention mechanism proposed in Reference [7] for ultra-high-speed weak target recognition. That study indicated that enhancing feature correlations across the spatio-temporal dimensions could effectively suppress motion blur interference in target discrimination. FocNet-SR's dual-channel decoupling module separates instrument semantics at the spatial level, while its spatio-temporal attention network enhances key frame features in the temporal dimension. Their synergistic operation achieved comparable feature retention in high-speed dynamic scenes, demonstrating that this approach to motion blur mitigation is not an isolated design but draws upon advanced dynamic feature modelling concepts from high-speed visual tasks. In the "smoke interference" scenario, FocNet-SR still did best. This was due to its multi-scale feature aggregation capabilities. In Figure 12(b), in the "bleeding occlusion" scenario, FocNet-SR reached a delay of 18.1 ms, better than MS-AST. Overall, FocNet-SR achieved both high recognition accuracy and low latency in real-time processing under actual surgical interference.

Finally, the study compared the performance differences of the three models in regard to rule compliance and edit distance. Compliance with standard procedures was primarily based on the Expert Consensus on Standardized Operating Procedures for Laparoscopic Cholecystectomy (2022 Edition), evaluated by three senior physicians across three dimensions: adherence to the standard sequence of steps, duration within recommended time windows, and compliance with critical maneuvers. Final compliance was defined as the "proportion of steps meeting the above specifications," with high-risk procedures (e.g., clamping, dissection) assigned a 1.5-fold weight to reflect their clinical significance. The relevant comparative outcomes are presented in Table 4.

As shown in Table 4, in terms of rule compliance, FocNet-SR achieved a threshold of 85.6% in the "severing" stage, which was the most technically challenging stage in the overall processing. Its average edit distance across the four stages was only 1.2, far lower than the comparison models. A lower edit distance indicated a closer approximation to the standard results, suggesting fewer errors and clearer data boundaries during processing. FocNet-SR's spatio-temporal attention mechanism and its strategy for dynamically segmenting the identified data jointly ensured the accuracy and stability of the final results. In summary, Table 3 fully validated FocNet-SR's high applicability in real-world clinical settings.

Table 4

Comparison of rule compliance and edit distance in critical surgical phases.

Phase	Metric	MS-AST	AMS-TCN	FocNet-SR
Clipping	Protocol compliance (%)	82.4	86.3	90.5
	Edit distance	2.8	2.0	1.2
Severing	Protocol compliance (%)	75.8	80.1	85.6
	Edit distance	3.2	2.5	1.5
Dissection	Protocol compliance (%)	85.1	88.5	92.3
	Edit distance	2.5	1.8	0.9
Hemostasis	Protocol compliance (%)	80.3	83.2	88.7
	Edit distance	3.0	2.2	1.1
Average	Protocol compliance (%)	80.9	84.5	89.3
	Edit distance	2.9	2.1	1.2

5. Summary and Future Work

To address the challenges of complex instrument-tissue interaction and unstable stage transitions in endoscopic surgical procedure recognition, this study proposed the FocNet-SR model. This model separated instrument and tissue features through a lightweight dual-channel feature decoupling module and introduced an improved multi-scale DeepLabV3+ and CBAM attention mechanism to focus on key regions. Experiments showed that the model achieved a frame-level accuracy of 95.8% and an inference speed of up to 57 fps, exhibiting high robustness and low latency in extreme surgical scenarios. The FocNet-SR model improved the accuracy, robustness, and clinical adaptability of surgical procedure recognition. However, current research still has certain limitations. Future work will study multimodal information fusion to build a more comprehensive and intelligent surgical cognition system. Furthermore, for extremely rapid phase transitions, the current fixed receptive field growth model of TCN exhibited limited temporal resolution. Drawing inspiration from the ultra-high-speed target recognition technique in Reference [7], future approaches could incorporate multi-granularity temporal attention to capture multi-scale temporal features, or design adaptive temporal receptive fields to dynamically adjust the modeling window. This would enhance sensitivity to rapid transition events while preserving long-term modeling capabilities.

Statement on Data Availability

The Cholec80 dataset used in this study is publicly available and can be obtained from the official website of the CAMMA research group: https://camma.unistra.fr/datasets. The dataset comprises 80 laparoscopic cholecystectomy videos, with each frame meticulously annotated according to seven standard surgical stages.

References

1. Abdani, S. R., Ariffin, S. M. Z. S. Z., Jamil, N., Ibrahim, S. 3D-Based Convolutional Neural Networks for Medical Image Segmentation: A Review. International Journal of Electrical and Computer Engineering Systems, 2025, 16(5), 347-363. https://doi.org/10.32985/ijeces.16.5.1
2. Aboudi, F., Drissi, C., Kraiem, T. A Hybrid Model for Ischemic Stroke Brain Segmentation from MRI Images Using CBAM and ResNet50-UNet. Int. J. Adv. Comput. Sci. Appl., 2024, 15(2), 950-962. https://doi.org/10.14569/IJACSA.2024.0150296
3. Chen, Y.-T., Ahmad, N., Aurangzeb, K. Enhancing 3D U-Net with Residual and Squeeze-and-Excitation Attention Mechanisms for Improved Brain Tumor Segmentation in Multimodal MRI. Comput. Model. Eng. Sci., 2025, 144(7), 1197-1224. https://doi.org/10.32604/cmes.2025.066580
4. Clero, K., Ed-Diny, S., Achalhi, M., Cherkaoui, M., El Harraki, I., El Fkihi, S., Benzakour, I., Soror, T., Rziki, S., Abdelali, H. A., Tagemouati, H., Bourzeix, F. Loosening Rocks Detection at Draa Sfar Deep Underground Mine in Morocco Using Infrared Thermal Imaging and Image Segmentation Models. Artif. Intell. Geosci., 2025, 6(1), 1-13. https://doi.org/10.1016/j.aiig.2025.100106
5. Dang, L., Li, S., Zhang, W., Hou, Y., Liu, Y. ETC-Net: An Efficient Collaborative Transformer and Convolutional Network Combining Edge Constraints for Medical Image Segmentation. Evolutionary Systems, 2025, 16(2), 1-20. https://doi.org/10.1007/s12530-025-09682-0
6. De, A., Wang, X., Zhang, Q., Wu, J., Cong, F. An Efficient Memory Reserving-and-Fading Strategy for Vector Quantization Based 3D Brain Segmentation and Tumor Extraction Using an Unsupervised Deep Learning Network. Cognitive Neurodynamics, 2024, 18(3), 1097-1118. https://doi.org/10.1007/s11571-023-09965-9
7. Fu, H., Xu, J., Li, Y., Zhou, H., Chen, X. Augmented Reality Based Navigation System for Endoscopic Transnasal Optic Canal Decompression. Journal of Shanghai Jiaotong University (Science), 2025, 30(1), 34-42. https://doi.org/10.1007/s12204-024-2722-3
8. Harouni, M., Goyal, V., Feldman, G., Michael, S., Voss, T. C. Deep Multi-Scale and Attention-Based Architectures for Semantic Segmentation in Biomedical Imaging. Comput. Mater. Contin., 2025, 85(10), 331-366. https://doi.org/10.32604/cmc.2025.067915

9. Jenifer, S., Mary, M. J. C. Convolutional Neural Networks for Medical Image Segmentation and Classification: A Review. Journal of Information Systems and Telecommunication, 2023, 11(4), 347-358. https://doi.org/10.61186/jist.37936.11.44.347

10. Jian, M., Wu, R., Chen, H., Fu, L., Yang, C. Dual-Branch-UNet: A Dual-Branch Convolutional Neural Network for Medical Image Segmentation. Computer Modeling in Engineering & Sciences, 2023, 137(1), 705-716. https://doi.org/10.32604/cmes.2023.027425

11. Komatsu, M., Kitaguchi, D., Yura, M., Takeshita, N., Yoshida, M., Yamaguchi, M., Kondo, H., Kinoshita, T., Ito, M. Automatic Surgical Phase Recognition-Based Skill Assessment in Laparoscopic Distal Gastrectomy Using Multicenter Videos. Gastric Cancer, 2024, 27(1), 187-196. https://doi.org/10.1007/s10120-023-01450-w

12. Konnai, Y., Fukumoto, K., Takeuchi, M., Takeuchi, R., Fujiwara, S., Yasumizu, Y., Tanaka, N., Takeda, T., Matsumoto, K., Kosaka, T., Kawakubo, H., Kitagawa, Y., Oya, M. Artificial-Intelligence-Based Surgical Phase Recognition in Robot-Assisted Radical Prostatectomy and Cross-Surgeon Validation. Annals of Surgical Oncology, 2026, 33(2), 1870-1877. https://doi.org/10.1245/s10434-025-18590-1

13. Lan, K., Gao, F., Jiang, X., Cheng, J., Fong, S. Enhanced Cutaneous Melanoma Segmentation in Dermoscopic Images Using a Dual U-Net Framework with Multi-Path Convolution Block Attention Module and SE-Res-Conv. Comput. Mater. Contin., 2025, 84(3), 4805-4824. https://doi.org/10.32604/cmc.2025.065864

14. Liu, S., Cheng, Y., Lin, Y., Jin, X., Wang, H., Wang, Y. A Cross-Scale Residual Attention Network with Convolutional Kolmogorov-Arnold for Medical Image Segmentation. Signal, Image and Video Processing, 2025, 19(14), 1-13. https://doi.org/10.1007/s11760-025-04822-0

15. Luo, F., Qin, Y., Li, M., Song, Q. A New Method for Segmentation of Medical Image Using Convolutional Neural Network. J. Opt. (India), 2024, 53(4), 3411-3420. https://doi.org/10.1007/s12596-023-01543-w

16. Ma, L.-Y., Qin, J.-H., Liu, Y.-B., Zeng, G.-F., Xu, B.-L., Huang, T.-T. MCA-DeepLabv3+: A Cupping Spot Image Segmentation Network Based on Improved DeepLabv3+. Signal, Image and Video Processing, 2025, 19(1), 1-9. https://doi.org/10.1007/s11760-024-03781-2

17. Miura, T., Abe, S., Manabe, T., Shibata, Y., Kosaka, T., Adachi, T. Efficient FPGA Implementation of a Convolutional Neural Network for Surgical Image Segmentation Focusing on Recursive Structure. Lecture Notes on Data Engineering and Communications Technologies, 2023, 176, 137-148. https://doi.org/10.1007/978-3-031-35734-3_14

18. Nagendram, S., Singh, A., Babu, G. H., Joshi, R., Pande, S. D., Ahammad, S. K., Dhabliya, D., Bisht, A., Aadarsh, A. Stochastic Gradient Descent Optimisation for Convolutional Neural Network for Medical Image Segmentation. Open Life Sciences, 2023, 18(1), 20220665. https://doi.org/10.1515/biol-2022-0665

19. Rhee, W., Lee, H. R., Chang, B.-S., Chang, S. Y., Kim, H., Rhee, H. Comparison of Deep Learning Models for Real-Time Neural Tissue Segmentation in Spinal Endoscopy. BMC Med. Imaging, 2025, 25(1), 470-470. https://doi.org/10.1186/s12880-025-01918-4

20. Song, X., Ge, H., Li, T. Remote Sensing Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+. Journal of Measurement Science and Instrumentation, 2025, 16(2), 205-215. https://doi.org/10.62756/jmsi.1674-8042.2025020

21. Tian, J., Zhou, Y., Chen, X., AlQahtani, S. A., Chen, H., Yang, B., Lu, S., Zheng, W. A Novel Self-Supervised Learning Network for Binocular Disparity Estimation. Computer Modeling in Engineering & Sciences, 2025, 142(1), 209-229. https://doi.org/10.32604/cmes.2024.057032

22. Waqas, M., Naseem, A. Artificial Intelligence in Sustainable Industrial Transformation: A Comparative Study of Industry 4.0 and Industry 5.0. FinTech and Sustainable Innovation, 2025, 1, Article A2. https://doi.org/10.47852/bonviewFSI52025321

23. Weng, D., Dou, S., Wang, H., Gong, D., Wang, Q., He, S. Infrared Image Segmentation Method Based on DeepLabV3+ for Identifying Key Components of Power Transmission Line. Progress in Electromagnetics Research C, 2023, 138(11), 191-203. https://doi.org/10.2528/PIERC23081905

24. Xiong, L., Hu, C., Wang, X., Wang, H., Tang, X., Wang, X. Detection and Threshold-Adaptive Segmentation of Farmland Residual Plastic Film Images Based on CBAM-DBNet. Int. J. Agric. Biol. Eng., 2024, 17(5), 231-238. https://doi.org/10.25165/j.ijabe.20241705.8069

25. Xue, W., Chen, C., Qi, X., Qin, J., Tang, Z., He, Y. M2ANet: Multi-Branch and Multi-Scale Attention

Network for Medical Image Segmentation. Chin. Phys. B, 2025, 34(8), 080703. https://doi.org/10.1088/1674-1056/adce96

26. Yang, S., Cai, Z., Luo, L., Ma, N., Xu, S., Chen, H. SurgPETL: Parameter-Efficient Image-to-Surgical-Video Transfer Learning for Surgical Phase Recognition. IEEE Transactions on Medical Imaging, 2026, 45(3), 1013-1023. https://doi.org/10.1109/TMI.2025.3615967
27. Zhang, C., Li, J., Wang, P.-L., Chen, H.-Y., Zhao, Y.-H., Wang, N., Zhang, Z.-T., Dang, Y.-W., Wang, H.-Q., Wang, J., Fu, C.-H. Multimodal Image Fusion-Assisted Endoscopic Evacuation of Spontaneous Intracerebral Hemorrhage. Chinese Journal of Traumatology, 2024, 27(6), 340-347. https://doi.org/10.1016/j.cjtee.2024.03.006
28. Zhang, X., Zhang, Q., Chen, J., Zhou, C., Wang, Y., Zhang, Z., Li, X., Qian, D. SPRMamba: Surgical Phase Recognition for Endoscopic Submucosal Dissection with Mamba. IEEE Sensors Journal, 2026, 26(6), 8887-8898. https://doi.org/10.1109/JSEN.2025.3649151
29. Zhang, Y., Yang, Q., Li, T., Zhang, F., Ren, Y., Li, Y., Xu, C. V-UNet: Medical Image Segmentation Based on Variational Attention Mechanism. CAAI Transactions on Intelligence Technology, 2025, 10(5), 1350-1362. https://doi.org/10.1049/cit2.70053
30. Zheng, L., Ge, Y., Zhu, M., Li, H., He, Y., Zheng, C. Fundus Image Segmentation Method Based on Improved DeepLabv3+ Network. International Journal of Wireless and Mobile Computing, 2025, 28(4), 416-426. https://doi.org/10.1504/IJWMC.2025.146642

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).