

Research Article

Image segmentation network for laparoscopic surgery

Kang Peng^{a,1}, Yaoyuan Chang^{b,1}, Guodong Lang^a, Jian Xu^b, Yongsheng Gao^{a,*},
Jiajun Yin^{b,*}, Jie Zhao^a

^a The State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150000, China

^b Zhongshan Hospital Affiliated to Dalian University, Dalian 116001, China

ARTICLE INFO

Article history:

Received 20 November 2024

Revised 16 March 2025

Accepted 21 April 2025

Available online 6 May 2025

Keywords:

Laparoscopic surgery image
Medical image segmentation
Convolutional neural networks
Attention mechanism
Feature fusion

ABSTRACT

Surgical image segmentation serves as the foundation for laparoscopic surgical navigation technology. The indistinct local features of biological tissues in laparoscopic image pose challenges for image segmentation. To address this issue, we develop an image segmentation network tailored for laparoscopic surgery. Firstly, we introduce the Mixed Attention Enhancement (MAE) module that sequentially conducts the Channel Attention Enhancement (CAE) module and the Global Feature Enhancement (GFE) module linked in series. The CAE module enhances the network's perception of prominent channels, allowing feature maps to exhibit clear local features. The GFE module is capable of extracting global features from both the height and width dimensions of images and integrating them into three-dimensional features. This enhancement improves the network's ability to capture global features, thereby facilitating the inference of regions with indistinct local features. Secondly, we propose the Multi-scale Feature Fusion (MFF) module. This module expands the feature map into various scales, further enlarging the network's receptive field and enhancing perception of features at multiple scales. In addition, we tested the proposed network on the EndoVis 2018 and a human minimally invasive liver resection image segmentation dataset, comparing it against six other advanced image segmentation networks. The comparative test results demonstrate that the proposed network achieves the most advanced performance on both datasets, proving its potential in improving surgical image segmentation outcome. The codes of MAMNet are available at: <https://github.com/Pang1234567/MAMNet>.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Image segmentation is the technical foundation for achieving the navigation of laparoscopic surgery [1,2]. It aims to apply algorithms to perform pixel-level classification of human organs, tissues, or other surgical instruments in laparoscopic image. During surgery, the results of laparoscopic image segmentation can assist surgeons in locating different organs, tissues, and lesions [3], reducing the workload on the surgeon and improving surgical efficiency and safety.

Laparoscopic image segmentation is more challenging compared to other medical image segmentation tasks, primarily due to the indistinct local visual features of biological tissues. The main reasons for this are twofold. First, there is a similarity between the local visual features of different organs and tissues [4]. Visual features in images can be primarily categorized into color features, texture features, shape features, and spatial relationship features. However, in laparoscopic images, different organs and

tissues often exhibit significant similarities in color and texture features, which result in indistinct local visual features. Moreover, real-time surgical images of laparoscopic surgery requiring a light source may be affected by the reflections of the surface of biological tissues and shadows resulting from obstructions between organs and instruments, which can make the local visual features of the biological tissues indistinct [5]. In Fig. 1, (a1) and (a2) are selected from the EndoVis 2018 dataset [6], while (b1) and (b2) are the corresponding ground truth of segmented images. Fig. 1 (a3) and (a4) are laparoscopic images of liver resection surgery provided jointly by Harbin Institute of Technology and Dalian University's Affiliated Zhongshan Hospital, with (b3) and (b4) being the corresponding ground truth of segmented images created by professional doctors. From Fig. 1 (b1)-(b4), it is clear that Regions I and II belong to different types of biological tissues. However, in actual surgical images, the similarity between biological tissues (Fig. 1(a1, a3)), reflections (a2), and shadows (a4) makes it very difficult to distinguish the local visual features of Regions I and II. In recent years, with the rise of intraoperative navigation technology, many researchers have focused on intraoperative medical image segmentation. Ni et al. [7] proposed RAUNet, which introduces a multi-level attention mechanism,

* Corresponding authors.

E-mail addresses: gaoy@hit.edu.cn (Y. Gao), yinjiajun@dlu.edu.cn (J. Yin).

¹ The two authors contributed equally to this work.

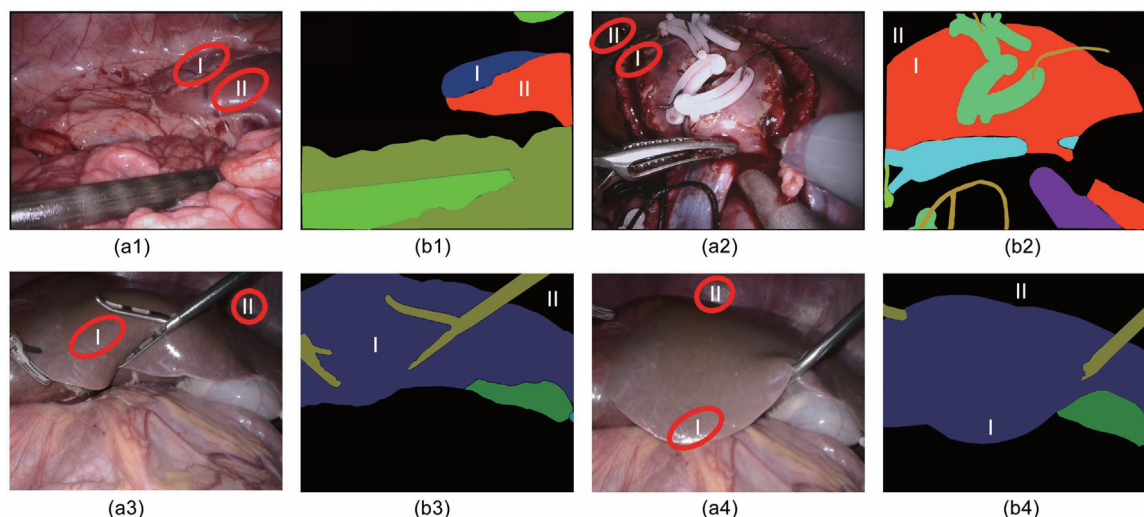


Fig. 1. (a1) and (a2) are actual laparoscopic images from kidney resection surgeries, with (b1) and (b2) being the corresponding ground truth of segmented image. (a3) and (a4) are actual laparoscopic images from liver resection surgeries, with (b3) and (b4) being the corresponding ground truth of segmented image.

enabling the network to capture important features at different levels and resolutions, thereby improving the segmentation accuracy of surgical instruments. BAANet [8] proposed two key modules: the Branch Balance Aggregation module and Block Attention Fusion module. The effectiveness of these two modules in improving the segmentation accuracy of surgical instruments was validated in the EndoVis 2017 and the EndoVis 2018 dataset, both of which consist of surgical laparoscopic images. However, much of the research focused on the problem of surgical instrument segmentation, neglecting the segmentation of biological tissues. In the laparoscopic environment, surgical instruments tend to have distinct features and high contrast against the biological tissue background, making them relatively easier to segment accurately.

The similarities in visual features increase the difficulty for these algorithms to achieve accurate segmentation of biological tissues in laparoscopic images. Addressing the issue of indistinct local features in biological tissues is crucial for improving the accuracy of laparoscopic image segmentation. SRBNet [4] improved accuracy in cataract surgery and laparoscopic kidney resection image segmentation by designing a space squeeze reasoning module and a low-rank bilinear fusion module. It is also one of the earliest studies to emphasize the issue of biological tissue segmentation in surgical images. Furthermore, due to the limited research on laparoscopic image segmentation, current laparoscopic image segmentation datasets that include biological tissue labels are much rarer compared to other medical image datasets. In 2018, the International Conference on Medical Image Computing and Computer Assisted Intervention launched a sub-challenge for robot scene segmentation and released a laparoscopic dataset for kidney resection image segmentation (EndoVis 2018 dataset) [6]. The EndoVis 2018 dataset, which includes segmentation labels for biological tissues in laparoscopic surgery, is the first international dataset available to the public. Therefore, enriching laparoscopic surgery image segmentation datasets and providing more segmentation labels for clinical laparoscopic surgeries are of great importance.

In this paper, we propose MAMNet, a novel laparoscopic image segmentation network that leverages hybrid attention enhancement and multi-scale feature fusion for improved performance. An encoder-decoder structure is utilized as the backbone. The MAE and MFF modules are designed to address the issue of indistinct visual features of biological tissues in laparoscopic images. The MAE module effectively captures global features within the

image, inferring indistinct local features from the global context. The MFF module obtains multi-scale features while suppressing noise, further enhances the network's ability to perceive global features and improves segmentation accuracy. Moreover, we conducted tests not only on the EndoVis 2018 dataset but also applied the proposed network to a dataset of human minimally invasive liver resection image segmentations. This dataset was jointly provided by Harbin Institute of Technology and Zhongshan Hospital affiliated with Dalian University, further demonstrating the broad applicability and clinical value of the algorithm.

In summary, the main contributions of this paper are as follows:

(1) To address the issue of indistinct visual features of biological tissues in laparoscopic images, the MAE module is introduced. This module extracts global feature information and infers local features from the global context.

(2) The MFF module is proposed to facilitate the fusion of semantic feature information at varying scales. This integration significantly enhances the network's ability to perceive global feature information.

(3) While the proposed network is tested on the EndoVis2018 dataset, it is also tested on unpunished dataset taken from a human minimally invasive liver resection operation, which was jointly provided by Harbin Institute of Technology and Zhongshan Hospital Affiliated to Dalian University. We compare six advanced image segmentation networks on both two datasets, and the test results stated that the proposed method demonstrated great performance on both datasets.

2. Related work

2.1. Surgical instrument segmentation

Laparoscopic surgery is a type of minimally invasive surgery. Currently, most work related to laparoscopic surgery image segmentation focuses on instrument segmentation in various surgical scenarios. Zhen et al. [7] designed RAUNet for cataract surgery instrument segmentation and created the first cataract surgery instrument dataset for semantic segmentation to test network performance. StereoScenNet [9] was the first to utilize left and right frame information from a stereoscopic surgical system to assist with instrument segmentation, improving the accuracy of instrument segmentation. Lee [10] designed a two-stage laparoscopic image segmentation method. Initially, the method

employs deep learning technique for coarse segmentation. Subsequently, it utilizes post-processing techniques such as erosion and GrabCut to refine and enhance the segmentation accuracy. U-NetPlus [11] modified the U-Net architecture by redesigning the encoder and decoder and evaluated it on the MICCAI 2017 EndoVis Challenge dataset, showing excellent performance in binary segmentation tasks. However, there remains scope for improvement in the accuracy of multi-class segmentation tests. Sun [12] designed a lightweight deep neural network that improved the real-time performance of the model while maintaining segmentation accuracy. To further enhance the accuracy of instrument segmentation, many recent studies have concentrated on expanding the network model's receptive field to obtain the contextual information in surgical images. DRR-Net [13] addressed the issues of insufficient local feature mapping and lack of temporal modeling information in instrument segmentation tasks by incorporating a novel non-connected attention recurrent convolutional block and a context fusion block. The network's effectiveness was demonstrated through testing on multiple instrument segmentation datasets. Yang [14] and colleagues proposed a multi-scale fusion network for endoscope image surgery instrument segmentation based on the transformer model. The study achieved the instrument segmentation accuracy through a dual-branch encoder structure, an attention feature fusion module, and an additive attention and concatenation module.

2.2. Attention mechanism

The attention mechanism [15] is a method that mimics the human visual and cognitive system; it can improve the performance and generalization capability of neural network and is widely applied in various computer vision tasks. Yang et al. [16] designed an attention-guided channel mechanism to enhance the segmentation between adjacent retinal layers in OCT images under the influence of choroidal neovascularization. CA-Net [17] used a comprehensive attention mechanism to enhance feature mapping between channels and multiple scales separately. Tests on datasets for skin lesion segmentation and fetal MRI multi-class segmentation demonstrated improvements in network performance. DANet [18] proposed a deformable attention network and integrated it into the U-Net architecture. Testing on a publicly available COVID-19 dataset demonstrated the effectiveness of this attention mechanism in improving lung infection segmentation performance. Yang et al. [19] proposed a dual attention module, enhancing encoder output feature maps with channel attention and decoder input feature maps with spatial attention. By connecting the feature maps generated by the two attention modules, the approach improved the precision of surgical instrument segmentation. CGBA-Net [20] proposed a bidirectional attention mechanism that integrates feature information from different directions using average pooling and maximum pooling operations, enhancing the accuracy of surgical instrument segmentation. HCTA-Net [21] proposed a multi-dimension attention module that rotates the input 3D feature map in three different directions and performs global pooling in each direction, obtaining attention weights from each direction. By multiplying these three attention weights, a comprehensive attention weight is generated, reducing background interference during segmentation to improve segmentation performance. Shokofeh Anari et al. [22] combined the pre-trained transformer model with the UNet framework to achieve breast tumor segmentation, demonstrating the potential of transformer's self-attention mechanism in improving medical image segmentation performance. Wang et al. [23] proposed a two-stage CNN method, which improved the accuracy of prostate MRI image segmentation. Zhuang et al. [24] developed a human body point cloud model for the

semantic segmentation of human body features. This work used LiDAR and an enhanced PointNet network with spatial feature extraction and channel attention modules to improve semantic segmentation accuracy by 4.5%, enabling precise identification of human body features for elderly bath care.

2.3. Feature fusion

In image segmentation, different levels of features capture different types of information. High-level features, having undergone multiple convolution operations, contain more semantic information but exhibit lower resolution and poor detail sensitivity. On the other hand, low-level features possess higher resolution and rich detail information but have lower semantic content and more noise. Therefore, many studies have attempted to process feature maps at different levels to obtain more comprehensive and precise feature information. Li et al. [25] proposed a Multi-scale Bottleneck Residual module, which addresses the challenge of balancing accuracy and model parameters in Scanning Laser Ophthalmoscopy images segmentation methods. PSPNet [26] is proposed by Zhao that utilized a pyramid pooling module and used dilated convolutions to enhance the performance and accuracy of neural network in image segmentation tasks. Zhang et al. [27] combined convolutional feature fusion, channel attention fusion, and encoder-decoder feature fusion to propose a comprehensive feature fusion strategy that improved the accuracy of vestibular segmentation in CT images. Ding et al. [28] proposed a multi-scale feature adaptive fusion method to enhance the performance of interactive image segmentation. Bendeche et al. [29] proposed an optimized Convolutional Neural Network (CNN) that used an Improved Chimp Optimization Algorithm (IChOA) to adjust all weight and bias values of the CNN model. This work adopted IChOA and SVM classifiers to select more key features and input these features into the optimized CNN model, improving the accuracy and robustness of the model for brain tumor segmentation. Yang et al. [30] introduced a hierarchical feature fusion method for achieving layered feature fusion. Zhai et al. [31] implemented a feature fusion method during up sampling to integrate low-level semantic features, thereby capturing more low-level detail information and enhancing the segmentation of retinal vessel images. Ranjbarzadeh et al. [31] proposed an automatic breast tumor segmentation and recognition based on a shallow CNN that uses multi-feature extraction routes. MFF-Net [32] designed a multi-channel feature map concatenation to enhance model generalization by fusing deep and shallow features. CFFR-Net [33] devised a feature fusion mechanism to coordinate global and local features extracted by different feature extraction modules. Du et al. [34] developed a cell image segmentation network model based on edge feature residual fusion. This model incorporates independent edge feature extraction and residual fusion modules to enhance edge features and constraints during cell target feature fusion, thereby improving the accuracy of cell contour localization. Bendeche et al. [33] proposed a framework for breast tumor segmentation and recognition in mammograms based on encoding method and a shallow cascade CNN. By generating multiple encoded images to extract unique features and remove the pectoral muscle, the framework utilizes a shallow network for pixel classification, avoiding the complexity of CNN. TBU-net [35] proposed a Fusion Layer module based on thresholding and logical operations, capable of integrating features from different regions. This approach addresses the challenge of accurately segmenting lesions with structural heterogeneity and blurred boundaries, such as melanoma, colon polyps, and breast cancer. Wang et al. [36] designed a road scene segmentation network based on multi-scale feature fusion and contextual information aggregation, using contextual information

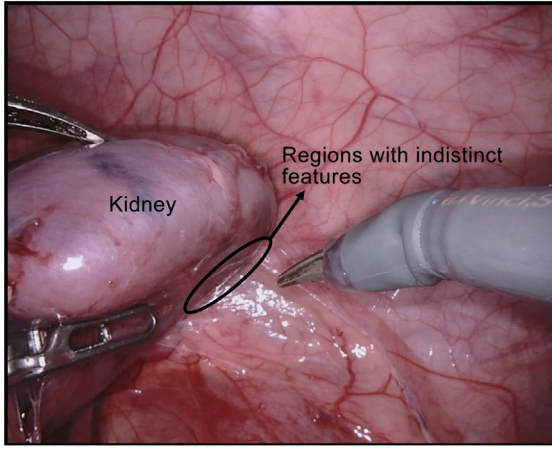


Fig. 2. Selected from the sample image of nephrectomy surgery. The area circled in black represents the junction of the kidney margin and the abdominal wall.

to guide feature fusion and enhancing the network's semantic feature extraction capability. Wang et al. [37] devised a multi-scale feature fusion method that continuously fuses low-level feature maps of different scales before the decoding, allowing the decoder to gradually learn the semantic information of the image. Feature fusion modules have been widely applied in image segmentation. In practical design, corresponding feature fusion methods are often tailored based on the characteristics and issues of specific segmentation tasks to improve segmentation accuracy.

3. Method

3.1. Overview

In laparoscopic images, it is hard to directly classify pixels based on the local visual features of indistinct biological tissues. The key to solving this issue involves first classifying regions with clear local features, and then inferring the boundaries of areas with unclear features based on global features such as contours and spatial relationships. As shown in Fig. 2, selected from the EndoVis 2018 dataset [6], it is clear that the area circled in black at the junction of the kidney margin and the abdominal wall has very similar color characteristics, making it difficult to directly determine the boundaries. However, inferring the boundaries of the less obvious areas relies on the kidney's global characteristics and spatial relationships. Therefore, how to expand the network's receptive field and better capture global semantic features to assist in the recognition of local features is key to solving the problem of laparoscopic image segmentation.

The proposed MAMNet structure, as shown in Fig. 3, utilizes Resnet34 [38] as the backbone for the encoder-decoder network. The MAE module is designed to enhance distinct local features and to obtain global features, then determining the boundaries of indistinct areas. To further expand the network's receptive field for global semantic features at different scales, we also proposed a Multi-scale Feature Fusion module, which integrates semantic features at different scales, thereby further improving the network's ability to capture global semantic features.

3.2. Mixed Attention Enhancement module

In order to enhance local features while simultaneously capturing shape features and spatial relationship features in different directions, we designed the MAE module, which consists of the Channel Attention Enhancement (CAE) module and the Global Feature Enhancement (GFE) module in series.

3.2.1. Channel Attention Enhancement module

The CAE module enhances the prominent channels of the original feature map after encoding. Since different channels correspond to different semantic features, this module effectively enhances the regions with prominent features in the laparoscopic image, providing a detailed input feature map for subsequent global semantic feature extraction. As illustrated in Eq. (1), performing global maximum pooling captures the prominent feature information in each channel and enhances the texture and color features in the image. Therefore, by employing a 1×1 convolution and a non-linear activation function in conjunction with global maximum pooling, the prominent feature information in each channel is encoded into a weight vector. Subsequently, the prominent feature weight is adaptively distributed to the input feature map using broadcast element-wise multiplication (Hadamard product), thereby enhancing the prominent features and filtering out irrelevant background features, as depicted in Eq. (2).

$$y_k^c = \max_{\substack{0 \leq i \leq H \\ 0 \leq j \leq W}} [\sigma(f(x_k(i, j)) + b)] \quad (1)$$

$$y = y_k^c \odot x \quad (2)$$

In Eq. (1) and (2), $x \in \mathbb{R}^{C \times H \times W}$ represents the input feature map, and $y \in \mathbb{R}^{C \times H \times W}$ represents the output feature map. k , i and j are the index values of the input feature map in the channel, height, and width dimensions, respectively. σ denotes the Linear rectification function, f represents the convolution operation with a 1×1 kernel, and b represents the bias term. The symbol \odot denotes the broadcast Hadamard product.

3.2.2. Global Feature Enhancement module

The GFE module boosts the network's ability to perceive global features, including overall shape and spatial relationships, thus enabling the inference of features in indistinct regions from distinct features. Previous research [39] has shown that while adjusting the size of the convolution kernel can partially increase the network's receptive field, this approach not only has limited effectiveness in enhancing network capabilities but also significantly raises the computational complexity of convolution operations. As shown in Eq. (3) and (4), the global feature enhancement module conducts global average pooling operations on the feature map separately from the height and width dimensions, compressing the overall features in both directions into two feature weights.

$$y_k^H = \frac{1}{H} \sum_{i=1}^H \delta_1 [f_1(x_k(i, j)) + b_1] \quad (3)$$

$$y_k^W = \frac{1}{W} \sum_{j=1}^W \delta_2 [f_2(x_k(i, j)) + b_2] \quad (4)$$

In Eq. (3) and (4), $x_k(i, j) \in \mathbb{R}^{C \times H \times W}$ represents the input feature map, y_k^H and y_k^W represent the feature weight in the height and width directions of the feature map, H and W represent the height and width of the feature map. δ_1 and δ_2 denote the Linear rectification function, f_1 and f_2 the convolution with a 3×3 kernel, b_1 and b_2 represent the bias term.

To integrate the global feature weights from two directions into a single three-dimensional feature weight matrix, we designed an extended Hadamard product guided by channel relationships. As shown in Fig. 4, after feature compression, the feature weights in the height direction have a size of $C \times H$, and the feature weights in the width direction have a size of $C \times W$. Due to the same channel relationships, any point $y(k, i$ and $j)$ in the three-dimensional feature weights should simultaneously

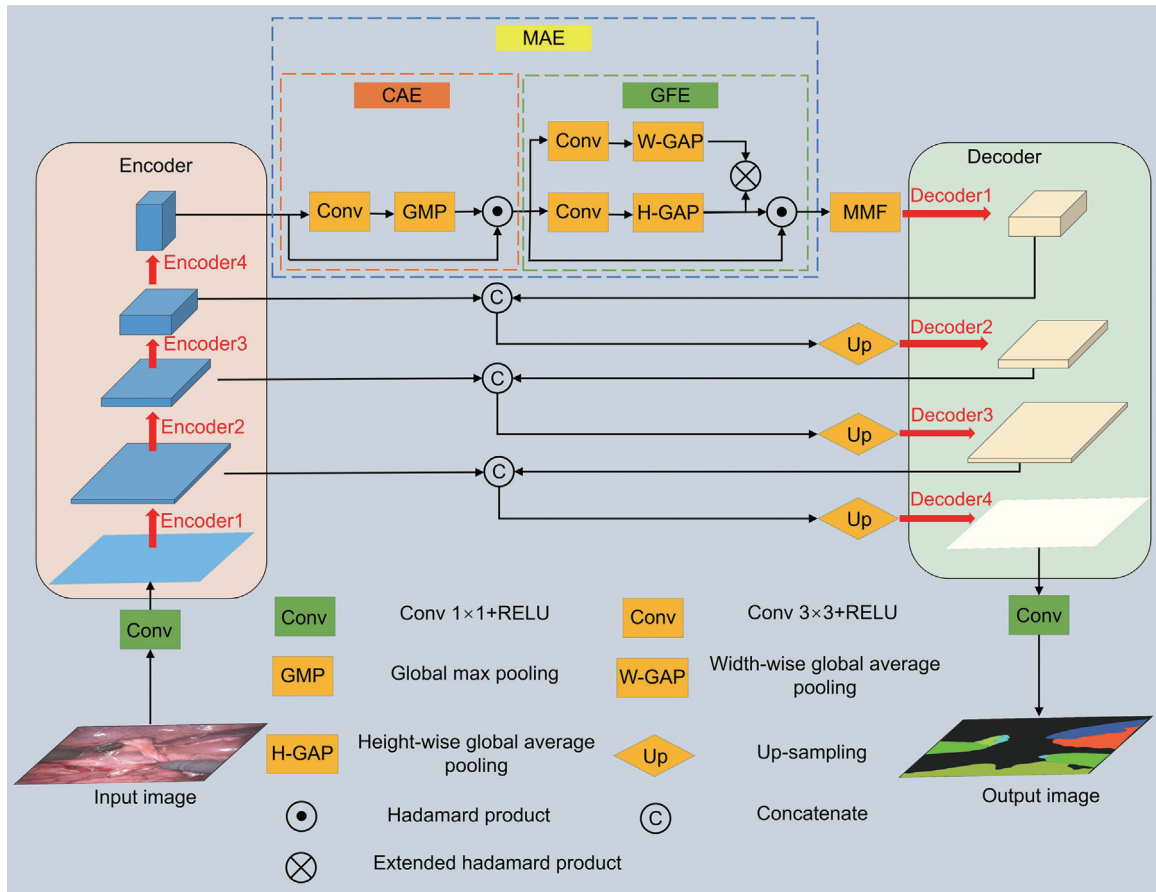


Fig. 3. Architecture of MAMNet.

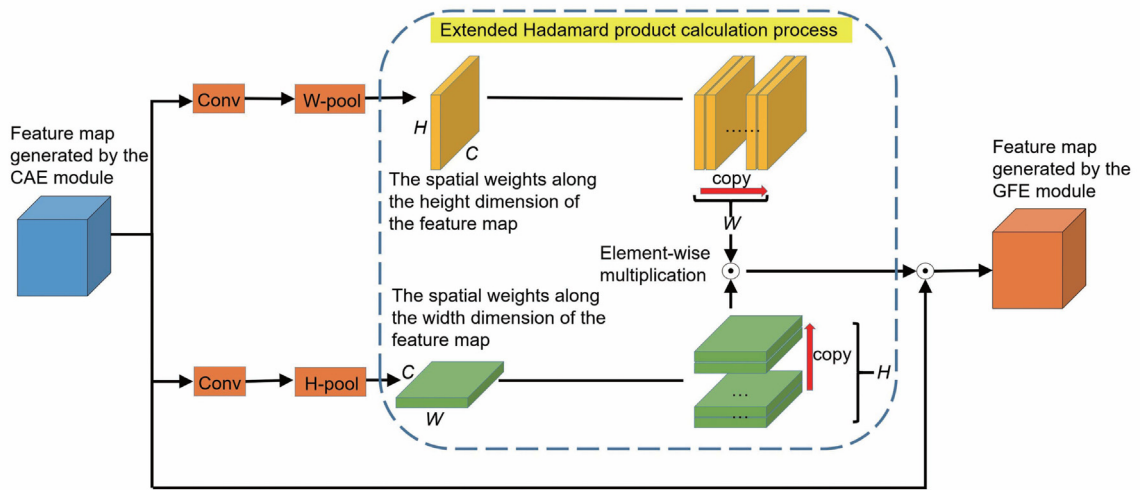


Fig. 4. Architecture of the GFE module. The extended Hadamard product calculation process corresponds to \otimes in Eq. (5).

represent the feature weights at index (k, j) in the height direction and index (k, i) in the width direction.

The extended Hadamard product calculation, presented in Eq. (5), fuses the feature weights from both the width and height directions and expands the dimensions to $C \times H \times W$. This operation results in a three-dimensional global feature weight of the same size as the input feature map, facilitating the subsequent feature fusion process.

$$y_k(i, j) = y_k^H(j) \times y_k^W(i) \quad (5)$$

where $y_k(i, j)$ represents the output feature map and (i, j) represents the index value in the width and height directions, respectively.

3.3. Multi-scale Feature Fusion module

To further enhance the network's capability to perceive global semantic features, the MFF module is designed to capture feature information at varying scales. The principle of the MFF module

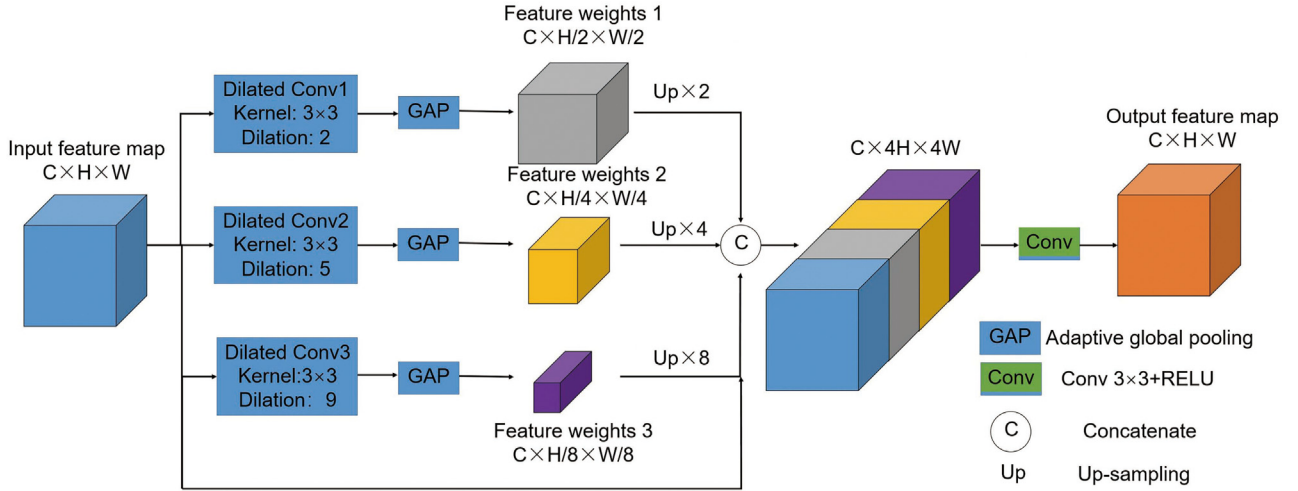


Fig. 5. Architecture of the MFF module.

is depicted in Fig. 5. As detailed in Eq. (6), for the input feature map with dimensions $C \times H \times W$, three distinct dilated convolution operations are initially performed. The kernels for these convolutions are all 3×3 , with dilation rates set to 2, 5, and 9, respectively. Dilated convolution allows for the expansion of the receptive field without resolution loss and preserves the relative spatial positions of pixels. When dilated convolutions with varying dilation rates are stacked, they can extract multi-scale information due to the differing receptive fields. Setting the dilation rates in a sawtooth pattern can effectively mitigate the grid effect [40]. Then we performed adaptive global pooling on the feature maps processed by dilated convolution, expanding them into feature weights of different sizes (feature weights 1–3), with the weight sizes being $C \times H/2 \times W/2$, $C \times H/4 \times W/4$, and $C \times H/8 \times W/8$ respectively. Adaptive global pooling can retain the global features of the image to the greatest extent.

$$y_k = \text{Pool}_{\text{avg}}^{(h_k, w_k)} (\sigma_k (f_k(x) + b_k)) \quad (6)$$

In Eq. (6), $x \in \mathbb{R}^{C \times H \times W}$ represents the input feature map, and $y_k \in \mathbb{R}^{C \times H \times W}$ represents the three-dimensional feature weight matrices of different sizes, where $k \in [1, 2, 3]$. $\text{Pool}_{\text{avg}}^{(h_k, w_k)}$ denotes the average pooling operation that compresses the feature map to height and width of (h_k, w_k) , respectively. f_k represents dilated convolution operations with a kernel of 3×3 and different dilated rates, b_k represents the bias term, and σ_k is the Linear rectification function. To ensure information consistency, these feature weight matrices are resized to match the dimensions of the input feature map $C \times H \times W$ through an up-sampling operation. Under the guidance of established channel relationships, we cascaded the input feature map with feature weights 1–3, and adaptively adjusted the cascaded feature maps to $C \times H \times W$ using a convolution operation with a 3×3 kernel. This allows features at different scales to be distributed adaptively across the entire feature maps. This multi-scale feature fusion and adjustment operation enables the extraction of global semantic information and reduces the interference of noise. As a result, the final network output delivers a more accurate segmentation image.

4. Experiments

Currently, among the publicly available laparoscopic segmentation international datasets, only the EndoVis 2018 dataset provides annotations for different types of biological tissue. The proposed MAMNet was initially validated on the EndoVis 2018

dataset, which comprises pig endoscopic images used as raw image data. These images represent a simpler environment compared to the human abdominal cavity. To further validate the broad applicability and clinical value of the algorithm, we conducted tests on the proposed network using a dataset of human minimally invasive liver resection images. For each dataset, we compared six advanced image segmentation networks. Due to variations in the precision of different dataset creations, to ensure the credibility of algorithm performance comparisons, all comparative experiments and our segmentation network tests employed the same method to divide the datasets. Finally, we conducted ablation studies on the MAE and MFF modules compared to six other networks to validate their practicality.

4.1. Dataset

4.1.1. EndoVis 2018 dataset

The dataset utilized in this study originates from the 2018 Robotic Scene Segmentation Challenge, comprising original endoscopic images of pigs captured using da Vinci X or Xi systems. It includes 2,235 images in the training set and 610 images in the test set, each with a resolution of 1280×1024 . This dataset provides annotations for eleven foreground object classes found in pig endoscopic images, encompassing eight types of surgical instruments and three types of biological tissues. To the best of our knowledge, this is the first publicly available dataset to include detailed annotations for complex biological tissues.

4.1.2. Human minimally invasive liver resection image segmentation dataset

In actual surgical scenarios, human tissue presents significantly greater complexity compared to pig tissue, thereby complicating the process of dataset annotation. This dataset was derived from real minimally invasive liver resection surgery videos, from which similar frame images were manually removed to ensure diversity. It comprises 1,165 training images and 388 validation images, each with a resolution of 1280×1024 , meticulously annotated by expert physicians. The annotations cover the liver surface, resected liver parenchyma, gallbladder, hemostatic clips, and surgical instruments, with the results stored as pixel-level segmentation masks. The annotation process underwent multiple reviews to ensure accuracy and consistency of the data labeling. The dataset is derived from the project titled “Research on Clinical Early Screening and Pathogenesis of Malignant Tumors in the Digestive System” (approved by the ethics committee,

Ethics Approval Number: KY2023-002-2). All data collection and processing strictly adhered to relevant ethical guidelines, with informed consent obtained from all participants.

4.2. Implementation details

MAMNet is implemented using the PyTorch framework. The network employs cross-entropy loss as the reference parameter for updating network weights. The learning rate is adjusted using an exponential decay strategy, starting from 0.0001. The Adam optimizer is utilized for training, with a batch size of 2 and an input image size of 1280×1024 . The total number of epochs is set to 300. For validation on the EndoVis 2018 Dataset, we randomly partitioned 2,235 images as the training set and 610 images as the test set. For validation on the on human minimally invasive liver resection image segmentation dataset, we randomly allocated 1,165 images as the training set and 388 images as the test set. To ensure the robustness of the results, we also performed five-fold cross-validation and averaged the results as the final performance metric. The primary hardware setup includes a NVIDIA GeForce RTX 3060 GPU and an Intel Core i7-12700 processor, with all training and testing conducted on this device.

During training, the dataset consists of meticulously curated and annotated images, resulting in high acquisition costs and limited feature variability. To address this, we applied data augmentation techniques such as random rotation, flipping, and Gaussian noise addition to enhance the diversity of the training data. Additionally, to mitigate overfitting, we incorporated dropout layers after the fully connected layers of the model.

To quantitatively assess the performance of the image segmentation network, the Dice Coefficient (Dice) and Intersection over Union (IoU) are commonly used evaluation metrics. These metrics measure the similarity between the ground truth and the predicted results, with higher values indicating better segmentation performance.

$$Dice = \frac{2 |G \cap P|}{|G| + |P|} \quad (7)$$

$$IoU = \frac{|G \cap P|}{|G \cup P|} \quad (8)$$

where G represents the ground truth, and P represents the predicted results.

Since the aim of this study is to segment different types of biological tissues and instruments in laparoscopic surgical images, the mean Dice (mDice) and mean IoU (mIoU) are used to evaluate the segmentation performance, which calculates the average Dice and IoU values across each segmentation object. We compared our proposed method with several state-of-the-art methods, including UNet, AUNet, RAUNet, DeepLabv3+, PSPNet and SRBNet. Below is a brief overview of these benchmark methods:

(1) UNet: Based on UNet, the Attention Mechanism is introduced to enhance the model's ability to focus on the region of interest, so as to improve the segmentation accuracy. A learning rate of up to 0.001 and a batch size of 2 are set in our experiment.

(2) AUNet: Based on UNet, the Attention Mechanism is introduced to enhance the model's ability to focus on the region of interest, so as to improve the segmentation accuracy. A learning rate of up to 0.001 and a batch size of 2 are set in our experiment.

(3) RAUNet: An improved model based on UNet, integrating Residual Connections and Attention Mechanisms. It demonstrates strong performance in surgical instrument image segmentation. A learning rate of up to 0.0005 and a batch size of 2 are set in our experiment.

(4) DeepLabv3+: Built on an encoder-decoder architecture, DeepLabv3+ incorporates Atrous Spatial Pyramid Pooling to capture multi-scale contextual information and enhance precise

Table 1

The average segmentation results tested on EndoVis 2018 dataset.

Method	Backbone	mIoU (%)	mDice (%)
UNet	Resnet34	53.12	60.81
AUNet	Resnet34	60.00	68.14
RAUNet	Resnet34	66.30	75.34
Deeplabv3+	Resnet34	73.74	80.93
PSPNet	Resnet34	75.85	82.40
SRBNet	Resnet34	75.57	82.24
MAMNet	Resnet34	78.19	83.84

The proposed MAMNet achieved the outstanding performance, with the highest score of mIoU and mDice being 2.34% and 1.44% higher, respectively, than the second highest method in this experiment.

boundary segmentation. In our experiment, three parallel atrous convolution layers are used, and each convolution layer has a dilated rate of 6, 12, and 18, respectively.

(5) PSPNet: PSPNet incorporates a Pyramid Pooling Module to capture multi-scale contextual information and enhance semantic segmentation performance. A learning rate of up to 0.0002 and a batch size of 2 are set in our experiment.

(6) SRBNet: Based on an encoder-decoder architecture, it integrates Space Squeeze Reasoning and Low-Rank Bilinear Feature Fusion techniques. It is the first method to be designed to address the segmentation of biological tissues in surgical images. A learning rate of up to 0.0001 and a batch size of 2 are set in our experiment.

To ensure the reliability of the comparison, all methods used for comparison adopt ResNet34 as the encoder-decoder backbone network.

4.3. Experiment results

4.3.1. Test results on two datasets

4.3.1.1. Test results on EndoVis 2018.

To evaluate the segmentation performance of MAMNet, we initially tested the network on the EndoVis 2018 dataset and compared its performance with six other segmentation networks under identical experimental conditions. The average test results from six other methods on the EndoVis 2018 dataset are presented in Table 1. MAMNet demonstrated the best performance compared to six other methods, achieving scores of 78% and 84% in mIoU and mDice, respectively, surpassing PSPNet—the second-best performing method—by 2.34% and 1.44%. SRBNet demonstrated excellent performance in surgical image segmentation and was the first network designed to address the problem of local feature similarity. Compared to SRBNet, MAMNet showed improvements of 2.62% and 1.60% on mIoU and mDice scores, respectively. Additionally, Deeplabv3+, a classic image segmentation network widely used in various medical image segmentation tasks, was also included in the tests. Compared to Deeplabv3+, MAMNet outperformed it by 4.45% and 2.91% on mIoU and mDice scores, respectively. Furthermore, compared to other classic medical image segmentation networks such as UNet, AUNet, and RAUNet, MAMNet showed an even greater advantage in both performance metrics.

To demonstrate MAMNet's superiority in addressing the issue of indistinct local features of biological tissues in laparoscopic images, the mIoU and mDice scores for each classification object tested on the EndoVis 2018 dataset are presented in Table 2. MAMNet demonstrates the best performance in the segmentation results for all categories except for the US-probe, and it is also very close to the optimal value for US-probe segmentation which is achieved by Deeplabv3+. To provide a more intuitive comparison of segmentation results of PSPNet, SRBNet and the proposed network, Fig. 6 aims to visually demonstrate which of

Table 2
The mIoU and mDice of each category tested on the EndoVis 2018 dataset.

Method	Index	Shaft	Clasper	Wrist	Parenchyma	Kidney	Thread	Clamps	Suction	Intestine	US-probe
UNet	mIoU	83.91	68.68	58.09	82.12	79.40	48.85	67.61	0.0	95.61	0.0
	mDice	91.24	81.43	73.49	90.18	88.52	65.64	80.68	0.0	97.80	0.0
AUNet	mIoU	84.68	66.03	60.61	91.86	91.79	60.38	71.17	0.0	95.16	38.32
	mDice	91.71	79.54	75.47	95.76	95.72	75.29	83.16	0.0	97.52	55.41
RAUNet	mIoU	89.34	74.53	70.44	93.51	92.51	58.33	61.73	56.67	85.90	36.34
	mDice	94.37	85.41	82.66	96.64	96.11	73.68	76.33	72.34	97.91	53.30
Deeplabv3+	mIoU	90.58	76.75	61.32	95.12	96.12	72.03	82.74	69.00	98.67	68.82
	mDice	95.06	86.84	76.03	97.50	98.02	83.74	90.55	81.66	99.33	81.53
PSPNet	mIoU	91.40	79.14	69.22	94.76	94.87	71.81	80.15	73.34	98.38	81.25
	mDice	95.51	88.35	81.81	97.31	97.37	83.59	88.98	84.62	99.18	89.65
SRBNet	mIoU	91.96	78.80	71.49	93.67	94.02	71.56	80.18	71.90	98.98	78.67
	mDice	95.81	88.14	83.38	96.73	96.91	83.43	89.00	83.65	99.49	88.06
MAMNet	mIoU	93.26	83.26	76.35	96.19	97.25	73.73	83.49	77.70	99.08	79.82
	mDice	96.51	90.87	86.59	98.06	98.60	84.88	91.00	87.45	99.54	88.78

The figure in bold indicates the highest results.

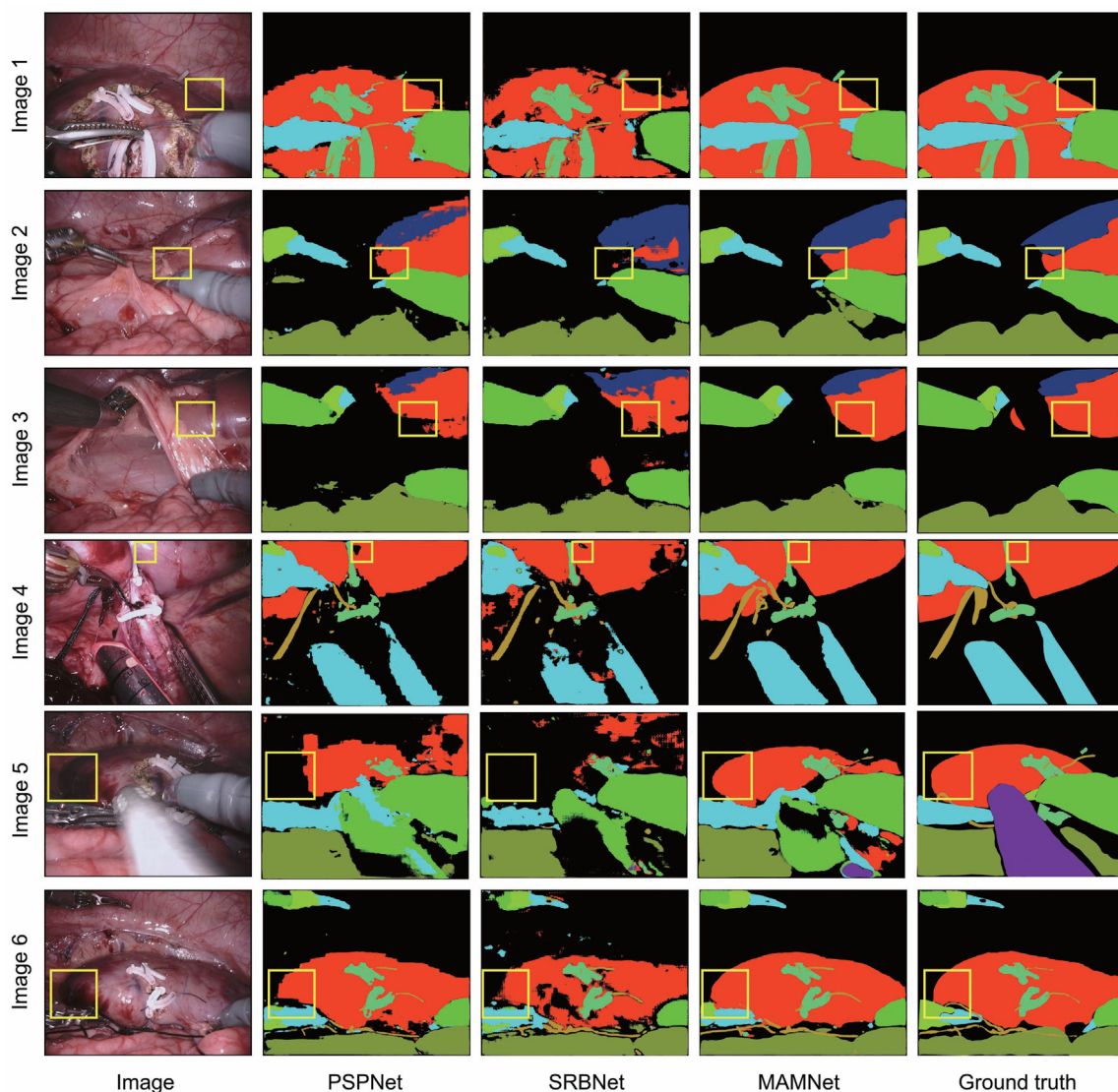


Fig. 6. Visualization of Segmentation Results by Different Methods: MAMNet, PSPNet, SRBNet, and ground truth image. Different objects are represented in distinct colors, with areas outlined in yellow indicating regions with indistinct local features. Compared to PSPNet and SRBNet, MAMNet achieves superior segmentation results with reduced noise.

the networks, PSPNet, SRBNet, or the proposed network, achieves segmentation results that are most similar to the ground truth and surgical images. In image 1, 2, and 3, the areas noted in

yellow are the regions of similar biological tissues. A comparative analysis with the ground truth reveals that PSPNet fails to accurately segment similar biological tissues, while SRBNet exhibits a

Table 3

The average segmentation results tested on the human minimally invasive liver resection image segmentation dataset.

Method	Backbone	mIoU (%)	mDice (%)
UNet	Resnet34	70.16	78.73
AUNet	Resnet34	73.69	81.69
RAUNet	Resnet34	73.89	83.83
Deeplabv3+	Resnet34	77.85	85.77
PSPNet	Resnet34	79.00	87.35
SRBNet	Resnet34	78.09	86.20
MAMNet	Resnet34	81.37	88.95

The proposed MAMNet achieved the outstanding performance, with the highest score of mIoU and mDice being 2.37% and 1.6% higher, respectively, than the second highest method in this experiment.

Table 4

The mIoU and mDice of each category tested on the EndoVis 2018 dataset.

Method	Index	Gallbladder	Instrument	Surface	Parenchyma	Clip
UNet	mIoU	80.46	86.33	90.60	75.08	18.35
	mDice	89.17	92.66	95.07	85.77	31.01
AUNet	mIoU	86.20	86.09	92.66	79.71	23.81
	mDice	92.58	92.53	96.19	88.71	38.46
RAUNet	mIoU	68.23	90.88	92.58	71.39	46.40
	mDice	81.11	95.22	96.14	83.31	63.39
Deeplabv3+	mIoU	89.41	89.22	92.86	79.75	38.02
	mDice	94.41	94.30	96.30	88.73	55.09
PSPNet	mIoU	87.69	87.21	92.53	77.70	49.83
	mDice	93.44	93.20	96.12	87.45	66.51
SRBNet	mIoU	88.80	87.92	92.89	79.53	41.30
	mDice	94.07	93.57	96.31	88.60	58.46
MAMNet	mIoU	92.95	88.17	93.14	78.06	54.54
	mDice	96.35	93.72	96.45	87.68	70.58

The figure in bold indicates the highest results.

certain level of improvement over PSPNet. However, the segmentation results of MAMNet are more complete and closer to the ground truth compared to SRBNet and PSPNet. In image 4, 5, and 6, the areas noted in yellow are the regions where local features are indistinct due to reflections and shadows. The results indicate that MAMNet also exhibits superior capability in addressing local similarity issues caused by reflection and shadow.

4.3.1.2. The test results on human minimally invasive liver resection image segmentation dataset.

To effectively validate the superiority of MAMNet's performance, the minimally invasive liver resection image segmentation dataset tests were conducted, which was created based on real-time liver resection surgery images. As shown in Table 3, MAMNet achieved an average mIoU score of 81.37% and an average mDice score of 88.95% on the minimally invasive liver resection image segmentation dataset. It exhibited the best performance compared to six other image segmentation networks, outperforming the second-ranking PSPNet by 2.37% and 1.6%, respectively. Table 4 presents the mIoU and mDice scores of seven segmentation networks tested for each classification object. MAMNet achieved the best results in most of the tested classification objects. Fig. 7 provides a visual comparison of the segmentation results of MAMNet, SRBNet, and PSPNet comparing with the ground truth.

4.3.2. Ablation study

This section conducts ablation experiments on the EndoVis 2018 dataset to verify the practicality of the two proposed modules, and the results are shown in Table 5.

4.3.2.1. Ablation study of the MAE module.

To validate the effectiveness of the MAE module, Model 1 to 4 were tested on the EndoVis 2018 dataset. Model 1, which adopts the most common encoder-decoder structure and does not include any feature processing module, showed the poorest

Table 5

The ablation study of the proposed network.

Method	MAE		MFF	mIoU(%)	mDice(%)
	CAE	GFE			
Model 1	×	×	×	54.79	62.99
Model 2	✓	×	×	67.23	76.52
Model 3	×	✓	×	71.13	79.03
Model 4	✓	✓	×	73.42	80.76
Model 5	×	×	✓	70.84	79.13
Model 6	✓	✓	✓	78.19	83.84

Model 6 is the proposed model in this paper(MAMNet).

performance with mIoU and mDice scores of 54.79% and 62.99%, respectively. Model 2, which includes the CAE module on top of the encoder-decoder structure, achieved mIoU and mDice scores of 67.23% and 76.52%, respectively. This Model emphasizes the local features of the segmentation target, enabling more precise segmentation in similar regions and thereby improving overall segmentation performance to a certain extent. Model 3 is an enhancement of Model 1, incorporating the GFE module to provide global feature perception capability, expanding the network's receptive field. The mIoU and mDice values for Model 3 are 71.13% and 79.03%, respectively. Model 4 simultaneously incorporates both the CAE and GFE modules, strengthening the network's perception of local salient features through the CAE module, while the GFE module facilitates global feature perception. When tested on the EndoVis 2018 dataset, Model 4 achieved mIoU and mDice values of 73.42% and 80.76%, respectively, higher than Model 2 and Model 3. Results from Model 1 to 5 indicate that both the CAE and GFE modules enhance the performance of network segmentation, with their combined use proving more effective than either module alone.

4.3.2.2. Ablation study of the MFF module.

To validate the effectiveness of the MFF module, Model 5 was tested on the EndoVis 2018 dataset. As indicated in Table 5, the MFF module is integrated with the encoder-decoder structure. The test results on the EndoVis 2018 dataset show that Model 5 achieved mIoU and mDice scores of 70.84% and 79.13%, respectively. These scores represent improvements of 16.05% and 16.14% over Model 1. These results demonstrate the MFF module's significant impact in enhancing the performance of the network.

Model 6, as presented in this paper, integrates the CAE, GFE, and MFF modules. By comparing the test results from Model 4 to 6, it becomes apparent that both the proposed the MAE and MFF modules significantly enhance the performance of the laparoscopic image segmentation network. Furthermore, the synergistic effect of combining these modules leads to superior performance than when using either module alone.

4.3.3. Statistical significance test

To rigorously evaluate the performance of our proposed method, we conducted a paired-samples t-test, comparing the mIoU scores of our proposed method with those of six baseline methods in pairwise fashion. We randomly selected five samples from the test set of the EndoVis 2018 Dataset, each containing 100 images, and tested both our proposed method and the comparative methods on the same samples. The null hypothesis (H_0) states the means of the two methods are equal (no significant difference), while the alternative hypothesis (H_1) states the means of the two methods are not equal (significant difference). We set the significance level at $\alpha = 0.05$. The test results are presented in Table 6.

The results of the paired samples t-test demonstrate that our method significantly outperforms all six baseline methods

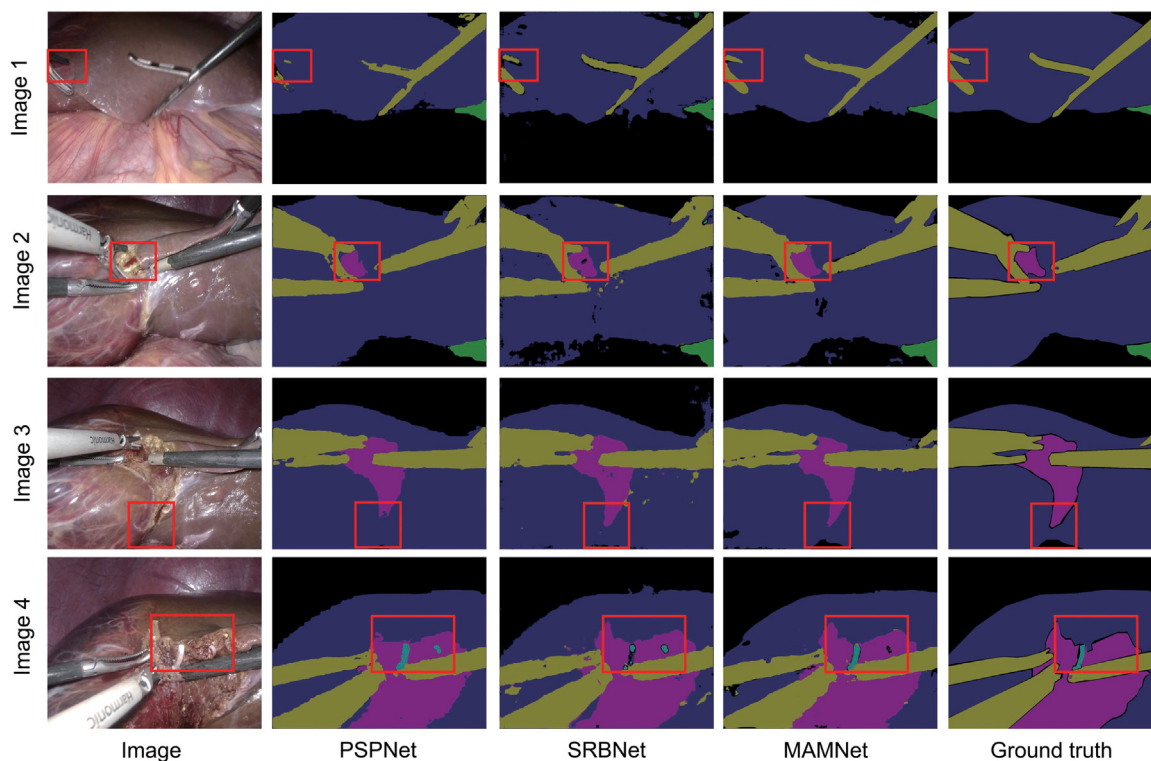


Fig. 7. Visualization of Segmentation Results by Different Methods: MAMNet, PSPNet, SRBNet, and ground truth image. Different objects are represented in distinct colors, with areas outlined in red indicating regions with indistinct local features. Compared to PSPNet and SRBNet, MAMNet achieves superior segmentation results with reduced noise.

Table 6
Statistical Significance of Performance Comparison (mIoU).

Comparison method	Mean difference (mIoU)	<i>t</i> -statistic	<i>p</i> -value	Significance
UNet	+24.55%	40.68	0.0000	Significant
AUNet	+17.66%	29.27	0.0000	Significant
RAUNet	+11.36%	18.83	0.0000	Significant
DeepLabv3+	+3.84%	4.92	0.0080	Significant
PSPNet	+1.40%	5.07	0.0071	Significant
SRBNet	+2.09%	5.11	0.0069	Significant

in terms of mIoU ($p < 0.05$ for all comparisons). This confirms the effectiveness of our approach in improving segmentation accuracy.

5. Discussion and conclusion

In this paper, we tackle the challenge of distinguishing indistinct local visual features in biological tissues during laparoscopic surgery image segmentation. A segmentation network that leverages the MAE and MFF modules is introduced. The MAE module, which is a series connection of the CAE module and GFE module, effectively captures global features and clarifies local features that are difficult to differentiate from global contexts. Additionally, we developed the MFF module, which is integrated into the decoder. This module enhances the fusion of feature information across various scales during the decoding process, thereby significantly improving the network's global feature perception. Compared to AUNet and RAUNet, the MAE module first enhances local features, and then extracts global semantic features to infer local details in regions with unclear visual features. This approach surpasses other algorithms in effectively integrating local and global features. Compared to DeepLabV3, PSPNet and SRBNet, our method combines the attention mechanism with multi-scale feature fusion method to extract more comprehensive feature information from multiple dimensions. It is worth noting that our proposed

extended Hadamard product for extracting global semantic features requires fewer parameters than the space extrusion module proposed by SRBNet.

We evaluated the proposed method on the EndoVis 2018 public dataset and compared it with six other medical image segmentation networks. The results confirmed that the proposed network outperforms these six segmentation networks. Notably, the EndoVis 2018 dataset is a collection of pig endoscope images, which are less complex than those of the human abdominal cavity. To enhance the reliability of the proposed network, we tested the network on a minimally invasive liver resection image segmentation dataset jointly provided by Harbin Institute of Technology and Zhongshan Hospital Affiliated to Dalian University. During the test on the EndoVis 2018 dataset, MAMNet demonstrated mIoU and mDice scores of 78.19% and 83.84%, respectively. On human minimally invasive liver resection image segmentation dataset, MAMNet demonstrated mIoU and mDice scores of 81.37% and 88.95%, respectively. In comparative tests with six other image segmentation networks, our proposed method demonstrated the highest performance based on mIoU and mDice across both datasets. Furthermore, the results of the statistical significance test showed that the differences between our method and the other six methods were all statistically significant ($p < 0.05$), indicating its application and clinical value. Finally, we

validated the effectiveness of the MAE module and MFF module through ablation experiments.

In future work, we can further improve the algorithm's performance in terms of generalization capability and computational efficiency through the following aspects:

(1) The segmentation model proposed in this paper is based on supervised learning, which has high demands on both the quality and quantity of the dataset. However, the datasets available for training must be created by specialist doctors, highlighting the limitations of the model's learning capability. Therefore, in future research, we will prioritize the development of semi-supervised and unsupervised algorithms for laparoscopic image segmentation to reduce the model's dependency on high-precision datasets.

(2) Taking liver resection as an example, there are significant variations in intra-abdominal fat accumulation and liver lesions among different patients during clinical surgery, leading to variations in the detailed features of laparoscopic images of the liver and surrounding areas. This diversity imposes higher demands on the model's generalization ability. To address this challenge, we plan to introduce transfer learning and expand the dataset by incorporating surgical images from different patients.

(3) In terms of computational efficiency, the MFF module may contain redundant weights, leading to the generation of unnecessary parameters and reducing overall efficiency. To address this issue, we need to refine the feature decomposition process to extract more useful features and apply network pruning. In future work, we will explore more refined pruning strategies to maximize computational efficiency while maintaining model accuracy.

In summary, while the proposed segmentation network has demonstrated promising results, its practical application is hindered by dataset dependency, limited generalization, and computational inefficiency. Future work will prioritize semi-supervised learning, transfer learning, and network pruning to enhance robustness and efficiency.

CRediT authorship contribution statement

Kang Peng: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Yaoyuan Chang:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis. **Guodong Lang:** Writing – review & editing, Validation, Software, Formal analysis. **Jian Xu:** Writing – review & editing, Validation, Formal analysis, Data curation. **Yongsheng Gao:** Validation, Supervision, Resources, Project administration, Funding acquisition. **Jiajun Yin:** Validation, Supervision, Resources, Project administration, Data curation. **Jie Zhao:** Supervision, Resources, Project administration, Data curation.

Ethics approval

This study was conducted in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical approval for the use of the minimally invasive liver resection dataset was obtained from the Ethics Committee of Zhongshan Hospital Affiliated to Dalian University (Approval No. KY2023-002-2, Date: April 3, 2024). Written informed consent was obtained from all individual participants included in the study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2022YFB4700700), the Dalian Deng Feng Program: key medical specialties in construction funded by the People's Government of Dalian Municipality, China ([2021]243), and the Liaoning Provincial Natural Science Foundation, China (2023JH2/101300102).

Data availability

The EndoVis 2018 Dataset originates from the 2018 Robotic Scene Segmentation Challenge and the human minimally invasive liver resection image segmentation dataset is confidential.

References

- [1] Nazim Haouchine, Jeremie Dequidt, Igor Peterlik, Erwan Kerrien, Marie-Odile Berger, Stéphane Cotin, Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery, in: 2013 IEEE International Symposium on Mixed and Augmented Reality, ISMAR, IEEE, 2013, pp. 199–208.
- [2] Toby Collins, Daniel Pizarro, Simone Gasparini, Nicolas Bourdel, Pauline Chauvet, Michel Canis, Lilian Calvet, Adrien Bartoli, Augmented reality guided laparoscopic surgery of the uterus, *IEEE Trans. Med. Imaging* 40 (1) (2020) 371–380.
- [3] Duygu Sarikaya, Jason J. Corso, Khurshid A. Guru, Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection, *IEEE Trans. Med. Imaging* 36 (7) (2017) 1542–1549.
- [4] Zhen-Liang Ni, Gui-Bin Bian, Zhen Li, Xiao-Hu Zhou, Rui-Qi Li, Zeng-Guang Hou, Space squeeze reasoning and low-rank bilinear feature fusion for surgical image segmentation, *IEEE J. Biomed. Heal. Inform.* 26 (7) (2022) 3209–3217.
- [5] Zhen Liang Ni, Pyramid attention aggregation network for semantic segmentation of surgical instruments, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI-20, 2020.
- [6] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al., 2018 robotic scene segmentation challenge, 2020, arXiv preprint arXiv:2001.11190.
- [7] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, Zhen Li, Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments, in: International Conference on Neural Information Processing, Springer, 2019, pp. 139–149.
- [8] Wenting Shen, Yaonan Wang, Min Liu, Jiazheng Wang, Renjie Ding, Zhe Zhang, Erik Meijering, Branch aggregation attention network for robotic surgical instrument segmentation, *IEEE Trans. Med. Imaging* (2023).
- [9] Ahmed Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, Øistein Hovde, Streoscennet: surgical stereo robotic scene segmentation, in: Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 10951, SPIE, 2019, pp. 174–182.
- [10] Eung-Joo Lee, William Plishker, Xinyang Liu, Timothy Kane, Shuvra S Bhattacharyya, Raj Shekhar, Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework, in: Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 10951, SPIE, 2019, pp. 461–469.
- [11] S.M. Kamrul Hasan, Cristian A. Linte, U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2019, pp. 7205–7211.
- [12] Yanwen Sun, Bo Pan, Yili Fu, Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery, *IEEE Robot. Autom. Lett.* 6 (2) (2021) 3870–3877.
- [13] Lei Yang, Yuge Gu, Guibin Bian, Yanhong Liu, DRR-Net: A dense-connected residual recurrent convolutional network for surgical instrument segmentation from endoscopic images, *IEEE Trans. Med. Robot. Bionics* 4 (3) (2022) 696–707.
- [14] Lei Yang, Yuge Gu, Guibin Bian, Yanhong Liu, TMF-Net: A transformer-based multiscale fusion network for surgical instrument segmentation from endoscopic images, *IEEE Trans. Instrum. Meas.* 72 (2022) 1–15.
- [15] Laurent Itti, Christof Koch, Ernst Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (11) (1998) 1254–1259.

- [16] Xiaoling Yang, Xinjian Chen, Dehui Xiang, Attention-guided channel to pixel convolution network for retinal layer segmentation with choroidal neovascularization, in: *Medical Imaging 2020: Image Processing*, vol. 11313, SPIE, 2020, pp. 786–792.
- [17] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, Shaoting Zhang, CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation, *IEEE Trans. Med. Imaging* 40 (2) (2020) 699–711.
- [18] Kumar Rajamani, Sahana D. Gowda, Vishwa Nedunoori Tej, Srividya Tirunellai Rajamani, Deformable attention (DANet) for semantic image segmentation, in: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2022*, pp. 3781–3784.
- [19] Lei Yang, Yuge Gu, Guibin Bian, Yanhong Liu, An attention-guided network for surgical instrument segmentation from endoscopic images, *Comput. Biol. Med.* 151 (2022) 106216.
- [20] Yiming Wang, Yan Hu, Junyong Shen, Xiaoqing Zhang, Heng Li, Zhongxi Qiu, Fangfu Ye, Jiang Liu, CGBA-Net: context-guided bidirectional attention network for surgical instrument segmentation, *Int. J. Comput. Assist. Radiol. Surg.* 18 (10) (2023) 1769–1781.
- [21] Lei Yang, Hongyong Wang, Guibin Bian, Yanhong Liu, HCTA-Net: A hybrid CNN-transformer attention network for surgical instrument segmentation, *IEEE Trans. Med. Robot. Bionics* (2023).
- [22] Shokofeh Anari, Gabriel Gomes De Oliveira, Ramin Ranjbarzadeh, Angela Maria Alves, Gabriel Caumo Vaz, Malika Bendechache, EfficientUNetViT: Efficient breast tumor segmentation utilizing UNet architecture and pretrained vision transformer, *Bioengineering (Basel)* 11 (9) (2024).
- [23] Zixuan Wang, Ruofan Wu, Yanran Xu, Yi Liu, Ruimei Chai, He Ma, A two-stage CNN method for MRI image segmentation of prostate with lesion, *Biomed. Signal Process. Control.* 82 (2023) 104610.
- [24] Chao Zhuang, Tianyi Ma, Bokai Xuan, Cheng Chang, Baichuan An, Minghuan Yin, Hao Sun, Deep learning-based semantic segmentation of human features in bath scrubbing robots, *Biomim. Intell. Robot.* 4 (1) (2024).
- [25] Peipei Li, Zhao Qiu, Yuefu Zhan, Huajing Chen, Sheng Yuan, Multi-scale bottleneck residual network for retinal vessel segmentation, *J. Med. Syst.* 47 (1) (2023) 102.
- [26] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, Pyramid scene parsing network, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017*, pp. 2881–2890.
- [27] Ruicong Zhang, Li Zhuo, Hui Zhang, Yan Zhang, Jinman Kim, Hongxia Yin, Pengfei Zhao, Zhenchang Wang, Vestibule segmentation from CT images with integration of multiple deep feature fusion strategies, *Comput. Med. Imaging Graph.* 89 (2021) 101872.
- [28] Zongyuan Ding, Tao Wang, Quansen Sun, Hongyuan Wang, Adaptive fusion with multi-scale features for interactive image segmentation, *Appl. Intell.* 51 (8) (2021) 5610–5621.
- [29] Ramin Ranjbarzadeh, Payam Zarbakhsh, Annalina Caputo, Erfan Babaei Tirkolaee, Malika Bendechache, Brain tumor segmentation based on optimized convolutional neural network and improved chimp optimization algorithm, *Comput. Biol. Med.* 168, 107723.
- [30] Dawei Yang, Yan Du, Hongli Yao, Liyan Bao, Image semantic segmentation with hierarchical feature fusion based on deep neural network, *Connect. Sci.* 34 (1) (2022) 1772–1784.
- [31] Zhengli Zhai, Shu Feng, Luyao Yao, Penghui Li, Retinal vessel image segmentation algorithm based on encoder-decoder structure, *Multimedia Tools Appl.* 81 (23) (2022) 33361–33373.
- [32] Zhenzhong Liu, Laiwang Zheng, Shubin Yang, Zichen Zhong, Guobin Zhang, MFF-Net: Multiscale feature fusion semantic segmentation network for intracranial surgical instruments, *Int. J. Med. Robot. Comput. Assist. Surg.* 20 (1) (2024) e2595.
- [33] Tahir Mahmood, Jin Seong Hong, Nadeem Ullah, Sung Jae Lee, Abdul Wahid, Kang Ryoung Park, CFFR-Net: A channel-wise features fusion and recalibration network for surgical instruments segmentation, *Eng. Appl. Artif. Intell.* 126 (2023) 107096.
- [34] Jinlian Du, Yanqiu Zhang, Xueyun Jin, Xiao Zhang, A cell image segmentation method based on edge feature residual fusion, *Methods* 219 (2023) 111–118.
- [35] LiFang Chen, Jiawei Li, Hongze Ge, TBUNet: A pure convolutional U-Net capable of multifaceted feature extraction for medical image segmentation, *J. Med. Syst.* 47 (1) (2023) 122.
- [36] Yunfeng Wang, Yi Zhou, Hao Wu, Xiyu Liu, Xiaodi Zhai, Kuizhi Sun, Chengliang Tian, Haixia Zhao, Tao Li, Wenguang Jia, et al., MFCANet: A road scene segmentation network based on multi-scale feature fusion and context information aggregation, *J. Vis. Commun. Image Represent.* 98 (2024) 104055.
- [37] Xun Wang, Xudong Zhang, Gan Wang, Ying Zhang, Xin Shi, Huanhuan Dai, Min Liu, Zixuan Wang, Xiangyu Meng, TransFusionNet: Semantic and spatial features fusion framework for liver tumor and vessel segmentation under JetsonTX2, *IEEE J. Biomed. Heal. Inform.* 27 (3) (2022) 1173–1184.
- [38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016*, pp. 770–778.
- [39] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015*, pp. 3431–3440.
- [40] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, Garrison Cottrell, Understanding convolution for semantic segmentation, in: *2018 IEEE Winter Conference on Applications of Computer Vision, WACV, Ieee, 2018*, pp. 1451–1460.