Medical Image Analysis 76 (2022) 102310

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

SurgiNet: Pyramid Attention Aggregation and Class-wise Self-Distillation for Surgical Instrument Segmentation

Zhen-Liang Ni[a,b], Xiao-Hu Zhou[a], Guan-An Wang[a,b], Wen-Qian Yue[a], Zhen Li[a], Gui-Bin Bian[a,b,*], Zeng-Guang Hou[a,b,c,*]

[a] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] The School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[c] Joint Laboratory of Intelligence Science and Technology, Institute of Systems Engineering, Macau University of Science and Technology, China

ARTICLE INFO

Article history:
Received 22 April 2021
Revised 1 October 2021
Accepted 22 November 2021
Available online 4 December 2021

Keywords:
Surgical Insturment Segmentation
Class-wise Self-Distillation
Pyramid Attention

ABSTRACT

Surgical instrument segmentation plays a promising role in robot-assisted surgery. However, illumination issues often appear in surgical scenes, altering the color and texture of surgical instruments. Changes in visual features make surgical instrument segmentation difficult. To address illumination issues, the SurgiNet is proposed to learn pyramid attention features. The double attention module is designed to capture the semantic dependencies between locations and channels. Based on semantic dependencies, the semantic features in the disturbed area can be inferred for addressing illumination issues. Pyramid attention is aggregated to capture multi-scale features and make predictions more accurate. To perform model compression, class-wise self-distillation is proposed to enhance the representation learning of the network, which performs feature distillation within the class to eliminate interference from other classes. Top-down and multi-stage knowledge distillation is designed to distill class probability maps. By inter-layer supervision, high-level probability maps are applied to calibrate the probability distribution of low-level probability maps. Since class-wise distillation enhances the self-learning of the network, the network can get excellent performance with a lightweight backbone. The proposed network achieves the state-of-the-art performance of 89.14% mIoU on CataIS with only 1.66 GFlops and 2.05 M parameters. It also takes first place on EndoVis 2017 with 66.30% mIoU.

1. Introduction

In recent years, intelligent perception has been promising in minimally invasive robotic surgery and computer-assisted microsurgery. Semantic segmentation of surgical instruments, whose goal is to segment instruments and identify corresponding categories, plays an essential role in assisted surgery (Sarikaya et al., 2017). The information, extracted from surgical instrument segmentation, can be used to navigate and control surgical robots. It also can be applied to provide real-time warnings for unnecessary and unsafe manipulation during surgery to improve surgical safety. Furthermore, semantic segmentation of surgical instruments offers numerous automated solutions for post-surgery work, such as objective assessment of surgical skills, surgical report generation, and surgical workflow optimization (Sarikaya et al., 2017). These applications can significantly reduce the workload of doctors.

Semantic segmentation of surgical instruments faces various challenges due to the complicated surgical scene. Intense light condition is essential for getting good visibility during surgery. However, it results in severe specular reflection which makes the appearance of the surgical instrument tend to be white. Besides, due to light angle change and biological tissue occlusion, there are shadows in several areas. The surgical instruments tend to be black in the shaded area, making it difficult to distinguish the surgical instruments from the background. The above problems can be summarized as illumination issues. These issues make surgical instrument segmentation challenging. Furthermore, to implement deployment on surgical robots, the network needs to be lightweight and occupy very few computing resources. However, a lightweight network is often difficult to obtain high-precision segmentation results. Neural networks often need to be deeper and complex to improve segmentation accuracy, which inevitably in-

* Corresponding authors.
E-mail addresses: guibin.bian@ia.ac.cn (G.-B. Bian), zengguang.hou@ia.ac.cn (Z.-G. Hou).

https://doi.org/10.1016/j.media.2021.102310
1361-8415/

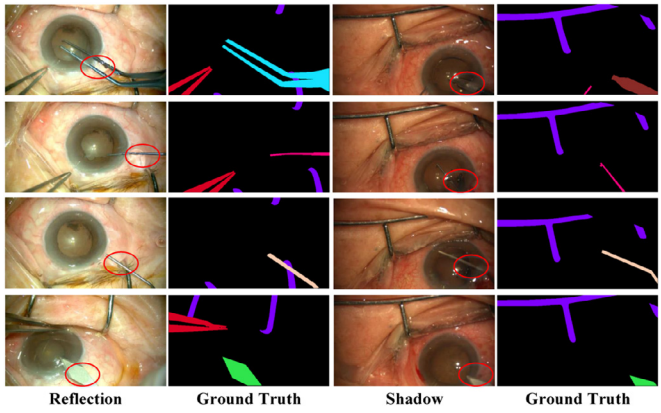

Fig. 1. Illumination issues in surgical instrument segmenation. The color and texture of instruments will change under different lighting. As we can see, the surgical instruments turns white under strong light but tends to be black under dark light.

creases computational costs. How to design a lightweight model to achieve high segmentation accuracy is still a very challenging topic.

Recently, several methods have been proposed for the semantic segmentation of surgical instruments. MF-TAPNet (Jin et al., 2019a) utilized optical flow as temporal clues to infer a prior indicating the location and shape of the instrument and improve the segmentation performance. ToolNet-C (Qin et al., 2019) combined the kinematic pose information to get the accurate silhouette mask. ST-MTL (Islam et al., 2021) enhanced the long-short term memory (LSTM) module to capture better long-term spatio-temporal dependencies for the surgical instrument segmentation and task-oriented saliency detection. A general embeddable approach (Qin et al., 2020) introduced the multi-angle feature aggregation (MAFA) method to adapt to instrument orientation variation and utilized the auxiliary contour supervision to make the contour more accurate. These methods introduce different prior knowledge and mechanisms to improve performance. However, they do not focus on illumination issues and do not consider the impact of high computational costs on deployment.

To deal with the illumination issue, the double attention module is designed to capture long-range semantic dependencies. The illumination issue changes the visual features of surgical instruments, including color and texture. Therefore, it is difficult for the network to identify surgical instruments based on these visual features directly. Considering that the visual features are spatially continuous, the features in the disturbed area can be inferred based on semantic dependencies. Specifically, the double attention module consists of two critical blocks: position attention block and channel attention block. The position attention block is based on low-rank bilinear pooling to model semantic dependencies between locations. The channel attention block encodes the semantic dependencies between channels by squeezing global information into an attention vector. The outputs of these two blocks are fused to generate attention features. Then, the attention features are calibrated to improve feature representation. The attention features model semantic dependencies between locations and channels. Thus, they can be used to infer semantic features in disturbed areas from adjacent pixels, dealing with illumination issues. Furthermore, the pyramid attention features are aggregated for final prediction. Local details in large-scale feature maps and overall shape features in small-scale feature maps can be captured, making predictions more accurate.

To compress the model while maintaining good performance, a class-wise self-distillation is proposed to enhance representation learning of the network. It is based on the class probability map generated by double attention features for knowledge transfer. The attention features can highlight the target area, so as to achieve a better distillation effect. Besides, the channel number of the distillation features is consistent with the total number of segmentation categories. Each channel reflects the feature distribution of a particular category. Thus, the feature distribution of each category can be learned separately by supervising specific channel features. In this way, students can better learn the feature distribution within the class without being disturbed by the feature distribution of other classes, helping to improve feature representation. Since high-level features contain richer semantic information, high-level probability maps are used as distillation targets and low-level probability maps are regarded as input. The input is encouraged to mimic the target by supervision. To adapt to the semantic segmentation task, the distillation operation is designed in the decoder, which helps to obtain more accurate high-resolution predictions. The class-wise self-distillation can significantly improve the learning ability of the network. Therefore, we can adopt a lightweight backbone while obtaining high accuracy.

According to the above analysis, the SurgiNet is proposed, which is based on the double attention module and class-wise self-distillation. The contributions of this work can be concluded as follows:

- The double attention module is designed to capture semantic dependencies between locations and channels. It can infer semantic features in the area disturbed by light, addressing illumination issues.
- Class-wise self-distillation is proposed to distill knowledge based on class probability maps, which can enhance the representation learning of the network. The network can adopt a lightweight backbone while obtaining high accuracy.
- The proposed SurgiNet achieves the state-of-the-art performance of 89.14% mIoU on CataIS with only 1.66 GFlops and 2.05 M parameters. It also takes first place on EndoVis 2017 with 66.30% mIoU.

2. Related Work

2.1. Surgical Instrument Segmentation

Recently, a series of methods have been proposed for the semantic segmentation of surgical instruments. Some work improved segmentation accuracy by capturing shape priors. For example, ToolNet-C combined with the kinematic pose information to get the accurate silhouette mask (Qin et al., 2019). MF-TAPNet (Jin et al., 2019b) adopted optical flow as temporal prior to provide a reliable indication of the instrument's location and shape for accurate segmentation. A general embeddable approach (Qin et al., 2020) utilized the auxiliary contour supervision to make the contour more accurate. Other work introduced various modules to improve feature representation. For instance, RAUNet (Ni et al., 2019) designed an attention module to fuse multi-level feature maps and emphasize the target region. A hybrid CNN-RNN method (Attia et al., 2017) adopted Recurrent Neural Network to expand the receptive fields, which also helps to capture global contexts. However, most of the methods have not focused on illumination issues, which limits their performance.

2.2. Attention Used in Semantic Segmentation

In recent years, the attention mechanism has been widely applied in semantic segmentation tasks. It helps to capture the global context and improve the feature representation. The attention mechanism can be divided into two categories according to the principle, namely position attention and channel attention. The

position attention is to model the semantic relationships between pixels, such as non-local block (Wang et al., 2018), A^2-Net (Chen et al., 2018b), and GCNet (Cao et al., 2019). Non-local calculated Embedded Gaussian of features to capture long-range dependencies. Pairwise Self-Attention Module (Wang et al., 2020a) applied subtraction to model the similarity of local features and used the Hadamard product to weight the similarity to the input features. The channel attention is to model the semantic dependencies between channels, which helps to emphasize target semantic features and improve feature representation. The squeeze-and-excitation network (Hu et al., 2018) is a typical representative. It squeezed global contexts into vector representation for modeling semantic dependencies between channels. Luminance-aware Pyramid Network (Li et al., 2020) proposed a multi-scale contrast feature block that used channel shuffle and channel scaling to capture the dependence between multi-scale feature channels. PAN (Li et al., 2018) captured the channel relationship in high-level features to guide low-level features and merged two-level features, which was based on the squeeze-and-excitation network (Hu et al., 2018). The above methods only capture one of spatial or channel attention. However, our double attention module captures both spatial and channel attention, which can better improve feature representation.

Besides, some work combines these two attention mechanisms for better performance, including dual attention network (Fu et al., 2019) and progressive attention guidance module (Zhang et al., 2018). DANet (Fu et al., 2019) adopted matrix multiplication to calculate the similarity of spatial and channel features respectively. And the attention features were projected back to the original feature space by matrix multiplication. Due to the matrix multiplication, their computational complexity is very high. Since the sample size of surgical instrument data sets is often small, DANet is prone to overfitting, resulting in poor generalization performance. Our double attention module uses low-rank bilinear pooling to approximate bilinear pooling, which uses the Hadamard product instead of matrix multiplication. Besides, global average pooling is used to capture dependencies between channels, which is also lightweight. Thus, the computational complexity of DAM is low, and it is not easy to produce over-fitting and more suitable for surgical instrument segmentation.

2.3. Distillation

Knowledge distillation is a key technology for model compression. It was proposed to transfer knowledge from a large model to a small model (Hou et al., 2019). Self-distillation aims to distill knowledge from the network itself, which is widely used in computer vision tasks. A self attention distillation (Hou et al., 2019) was proposed for lane detection, which distilled knowledge based on the attention map generated by different levels of features. Self-distillation was also applied to enhance tiny tissue segmentation (Zhou et al., 2020). Faster ReID (Wang et al., 2020b) adopted probability distillation to transfer knowledge from long codes to short codes within a network and utilized similarity distillation to transfer knowledge from a large model to a small model. (Yun et al., 2020) learned class-wise knowledge by distilling two batches of samples with the same label. It only distilled the classification results and does not distill the features between layers. Therefore, it has no way to realize the knowledge transfer between layers. (Wang et al., 2020c) proposed a knowledge distillation to transfer the knowledge from the large model to the small model. However, due to the small surgical scene data set, too large a model will cause over-fitting. Therefore, we design a class-wise self-distillation method to realize the transfer of knowledge from deep to shallow, which does not require training a large teacher model. (Zhang and Sabuncu, 2020) provided evidence that diversity in teacher predictions was correlated with the performance of self-distillation, and proposed an instance-specific regularization method to promote predictive diversity. The main contribution of this work is to propose a new regularization method to improve classification accuracy. However, it does not propose innovative distillation forms and does not improve the selection of distillation features. (Zhang et al., 2019) distilled the hidden layer features and fully connected layer from deep to shallow for classification tasks. Different from the above methods, our method distills the class probability map generated by double attention features. The double attention features can highlight the target area, helping to achieve a better distillation effect. Besides, our class-wise self-distillation performs feature distillation within the class to eliminate interference from other classes. And to adapt to the segmentation task, distillation is set in the decoder to obtain a more accurate high-resolution mask.

3. Methodology

3.1. Overview of Network Architecture

The SurgiNet is proposed for surgical instrument segmentation, which consists of pyramid attention aggregation and class-wise self-distillation. To address illumination issues, the double attention module (DAM) is proposed to capture both semantic dependencies between locations and channels. Based on the semantic dependence, it can infer the semantic features in the disturbed areas. Besides, to compress the model while maintaining good performance, class-wise self-distillation (CSD) is proposed to enhance the representation learning of the network. It performs feature distillation within the class to eliminate interference from other classes.

The architecture of the proposed SurgiNet is shown in Fig. 2. The MobileNetV2 (Sandler et al., 2018) is adopted as the backbone. Pyramid attention features are aggregated for the final prediction, helping to capture multi-scale features. The aggregated attention features are 1/4 of the input size, which helps to reduce the computational cost. They are upsampled four times to obtain a high-resolution mask.

3.2. Double Attention Module

The double attention module contains two attention blocks: position attention block and channel attention block. They capture semantic dependencies between locations and channels, respectively. The output of these two blocks is fused and calibrated. The architecture of the double attention module is shown in Fig. 3.

3.2.1. Position Attention Block

Due to illumination issues such as specular reflections or shadows, the color and texture of surgical instruments are changed. Therefore, it is difficult for the network to locate surgical instruments based on these features. To deal with this issue, we design the position attention block which is based on low-rank bilinear pooling to capture semantic dependencies between pixels.

A series of work uses bilinear pooling to capture attention features (Chen et al., 2018b; Kim et al., 2018). The low-rank bilinear pooling uses matrix decomposition and Hadamard product to approximate the bilinear pooling (Kim et al., 2016; Yu et al., 2017). Bilinear pooling can capture second-order statistics and encode more complex semantic dependencies. The low-rank bilinear pooling retains the ability to model complex semantic dependencies. Besides, it avoids matrix multiplication to reduce computational costs. Specifically, it is described in Eq. (1).

$$z = P^T (U^T x \otimes V^T y) + b \tag{1}$$

where $x \in R^N$ and $y \in R^M$ refer to the input vector. $z \in R^C$ refers to the output vector. $\otimes$ denotes Hadamard product. $U \in R^{N \times k}$ and

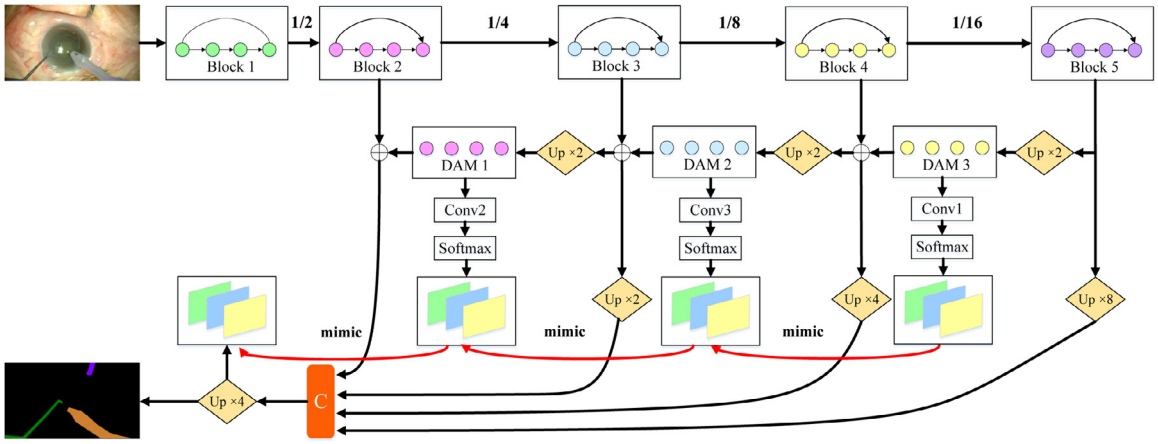

Fig. 2. The architecture of the SurgiNet. Double attention module (DAM) captures semantic dependencies between locations and channels to infer semantic features in disturbed areas. Besides, class-wise self-distillation (CSD) distills knowledge from multi-level class probability maps to enhance the representation learning of the network. Due to the enhancement of self-learning, the network can get high performance with little computational cost, which facilitates the deployment.

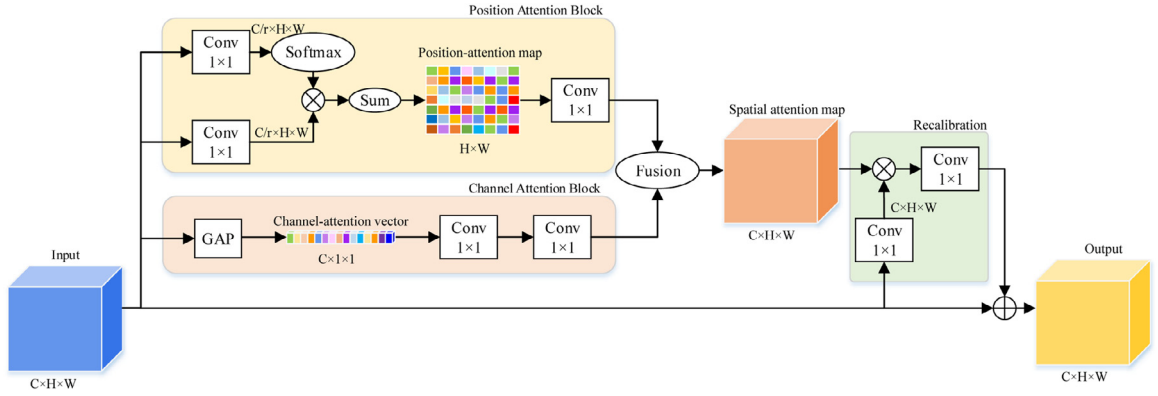

Fig. 3. The architecture of the double attention module. It contains the position attention block and the channel attention block. Their outputs are fused and calibrated, generating attention features. $\otimes$ represents the Hadamard product, and $\oplus$ denotes the addition.

$V \in R^{M\times k}$ are linear projections. $P \in R^{k\times C}$ is used to control the dimensions of the output. $b \in R^{C}$ is the bias.

The position attention block is designed based on a variant of low-rank bilinear pooling, which is described in Eq. (2). 1×1 convolutions are adopted to replace matrices U and V for linear projection. Non-linear activations help to improve feature representation (Kim et al., 2016). Thus, we adopt ReLU to add non-linear activations. Besides, softmax is adopted to normalize the feature map. Sum pooling corresponds to the matrix P in Eq. (1), which is utilized to adjust the dimension of the output. Finally, 1×1 convolution is applied to transform the position-attention map and improve feature representation.

The architecture of the position attention block is shown in Fig. 3. In addition to linear projection, 1×1 convolutions are also used to reduce the dimension of input feature maps for reducing computational costs. The dimension is reduced to C/r, where C is the original dimension. In this paper, r is set to 2. The input feature maps are $x \in R^{C\times H\times W}$ and $y \in R^{C\times H\times W}$, where H and W refer to the height and the width of the input respectively. The position-attention map is represented as $A_p \in R^{H\times W}$.

$$A_p = g[\delta(Ux) \otimes \sigma[\delta(Vy)]] + b \tag{2}$$

where $\otimes$ denotes Hadamard product. U and V denote 1×1 convolutions. δ denotes ReLU function. σ refers to softmax function and g represents the sum pooling.

Sum pooling and Hadamard products both contribute to reducing the computational cost. Thus, the position attention block is lightweight, which helps its deployment on surgical robots.

3.2.2. Channel Attention Block

The feature representation of the convolution operation is local. The network cannot capture long-range semantic dependencies, which results in a poor semantic understanding. Besides, different semantic features have different responses in various channels. Thus, target semantic features can be highlighted based on the semantic dependencies between channels. According to the above analysis, we design the channel attention block to capture global contexts and model semantic dependencies between channels.

Channel attention block adopts the global average pooling to capture attention features. The global average pooling squeezes input feature maps into an attention vector. This vector encodes the semantic dependencies between channels. Furthermore, each element of the attention vector aggregates global contexts, which helps to capture long-range semantic dependencies. The global average pooling is shown in Eq. (3).

$$a_k = \varphi(x_k) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_k(i,j) \tag{3}$$

where φ represents the global average pooling. $x_k \in R^{H\times W}$ refers to input feature maps. $k = 1, 2, ..., C$. The channel-attention vector is represented as $A_c = [a_1, a_2, ..., a_C]$.

Finally, two 1×1 convolutions with batch normalization are used to transform the channel-attention vector, helping to refine the semantic relationships.

3.2.3. Fusion of Attention Features

The position-attention map and channel-attention vector are fused before calibration. Specifically, we perform the broadcast Hadamard product on the position-attention map and channel-attention vector. The fused result has the same size as the input feature map.

$$A_s(k,h,w) = A_p(h,w) \odot A_c(k) \tag{4}$$

where A_s represents the spatial attention map. $k = 1, 2, ..., C$, $h = 1, 2, ..., H$ and $w = 1, 2, ..., W$. $\odot$ denotes broadcast Hadamard product.

3.2.4. Recalibration

We calibrate the spatial attention map A_s to further improve the feature representation. The calibration consists of semantic feature calibration and nonlinear transformation. First, we utilize the semantic information in transformed input features to calibrate A_s, as shown in Eq. (5). 1×1 convolution with Batch Normalization and ReLU is applied to adjust input features. The parameters of the convolution will be continuously updated to adapt to the distribution of the data by training. Thus the semantic information in input features is continuously optimized and calibrated. Besides, Batch Normalization is applied after convolution to normalize input features. ReLU is used to make features greater than 0 and increase non-linearity. These operations can improve feature representation and help calibrate attention features.

$$\widehat{A_s} = A_s \otimes \delta(f(x,\theta)) \tag{5}$$

where $\otimes$ denotes Hadamard product. x represents the input feature map. f refers to the 1×1 convolution. θ represents parameters of convolution. δ denotes ReLU function.

Then, we perform 1×1 convolution with Batch Normalization and ReLU to transform the spatial attention map. These operations help to calibrate attention features. It is worth noting that calibration is crucial. It can significantly improve performance in the experiment.

3.3. Class-wise Self-Distillation

To compress the model while maintaining good performance, the class-wise self-distillation (CSD) is proposed to reinforce representation learning of the network. Its goal is to perform top-down and multi-stage class probability distribution distillation. Class-wise self-distillation is based on the class probability map generated by double attention features for knowledge transfer. The attention features can highlight the target area, helping to achieve a better distillation effect. Besides, the channel number of the distillation features is consistent with the number of segmentation categories. Each channel reflects the feature distribution of a particular category. Thus, the feature distribution of each category can be learned separately by supervising specific channel features. In this way, students can better learn the feature distribution within the class without being disturbed by the feature distribution of other categories, improving the segmentation accuracy. To adapt to the semantic segmentation task, the distillation operation is designed in the decoder, which helps to obtain more accurate high-resolution predictions.

Specifically, class probability maps used for distillation are generated from double attention features. 1×1 convolution is applied to adjust the number of channels for attention features to the number of segmentation classes. After that, softmax is utilized to generate class-wise probability maps. Each channel reflects the probability distribution of a specific category. By supervising specific channel features, the feature distribution of the corresponding category is spread within the category.

$$\Phi(A_n) = \sigma\left(\frac{f(A_n,\theta)}{T}\right) \tag{6}$$

where A_n represents the features of the $n-th$ layer for distiilation. f refers to 1×1 convolution for probability map generation. θ represents the parameters of convolution. T is a temperature hyperparameter, which is set to 1 empirically. σ is the softmax function.

Since high-level features contain richer semantic information, high-level probability maps are used as distillation targets and low-level probability maps are regarded as inputs. The input is encouraged to mimic the targets by supervision. To ensure that the size of the target is consistent with the input, pooling is performed on the high-level probability map. Thus, the target generation can be expressed as the following equation.

$$\Psi(A_n) = \sigma\left(\frac{\beta(f(A_n,\theta))}{T}\right) \tag{7}$$

where $\beta(.)$ refers to pooling operation. T is also set to 1.

L_2 loss is utilized to evaluate the similarity of the probability map.

$$L_{distill} = \sum_{n=1}^{N-1} L_2(\Phi(A_n), \Psi(A_{n+1})) \tag{8}$$

where the $\Psi(A_{n+1})$ is the target of the distillation loss. N is the number of layers. As shown in Fig. 2, $N = 4$.

The total loss consists of segmentation loss and distillation loss, which is shown in Eq. (9). Focal loss (Lin et al., 2017b) is adopted as segmentation loss, which helps to deal with the class imbalance issue.

$$L = L_{seg}(p,g) + \alpha L_{distill} \tag{9}$$

where L_{seg} indicates the focal loss. p refers to prediction and g refers to ground truth. α is used to balance the weight of segmentation loss and distillation loss.

CSD is only adopted during the training phase. In the testing phase, the distillation branch will be removed from SurgiNet, which includes convolution and softmax used to generate the class probability map. Thus, CSD brings no computational cost in the deployment. And it does not require any additional labels.

4. Experiments And Results

The proposed SurgiNet is evaluated on the CataIS and the EndoVis 2017. It achieves the state-of-the-art performance of 89.14% mIoU on CataIS with only 1.66 GFlops and 2.05 M parameters and takes first place on EndoVis 2017 with 66.30% mIoU.

4.1. Dataset

4.1.1. CataIS

CataIS is a cataract surgical instrument dataset for semantic segmentation. It is extended from the Cata7 dataset (Ni et al., 2019) constructed by us. Two video sequences with darker lighting are added, and a new category is labeled. The CataIS contains 9 videos with a total of 2671 images. 850 images (video 7,8,9) are used for the test and 1821 images (video 1,2,3,4,5,6) are utilized for training. The resolution of images is 1920×1080. To speed up training and testing, each image in CataIS is resized to 640×384 pixels. There are 11 types of surgical instruments in this dataset. The quantity of each surgical instrument is shown in Table 1.

Table 1
The description for the CataIS dataset. CataIS contains a total of 11 surgical instruments. There are a total of 2671 frames, of which 1821 are used for training and 850 are used for test.

Instrument	Train	Test	Total
Primary Incision Knife (I1)	35	21	56
Secondary Incision Knife (I2)	42	25	67
Viscoelastic Cannula (I3)	277	148	425
Capsulorhexis Forceps (I4)	213	116	329
Micromanipulator (I5)	525	169	694
Lens Hook (I6)	388	71	459
Aspiration Handpiece (I7)	414	224	638
Implant Injector (I8)	57	45	102
Phacoemulsifier Handpiece (I9)	534	197	731
Bonn Forceps (I10)	238	171	409
Eyelid Retractors (I11)	1807	850	2657
Total	4530	2037	6567
Number of Frames	1821	850	2671

4.1.2. EndoVis 2017

EndoVis 2017 is from the MICCAI Endovis Challenge 2017 (Allan et al., 2019), which is based on endoscopic surgery. This dataset is acquired from a DaVinci Xi robot. It contains 3000 images with a resolution of 1920×1080, including 1800 images for training and 1200 images for the test. The test set contains 10 video sequences, and each sequence is used as a sub-dataset. We follow the training rules of EndoVis2017. When evaluating one of the first 8 test sequences, we use training data except the corresponding training sequences for training. When evaluating the last 2 test sequences, we use all the training data for training. There are 7 types of surgical instruments in EndoVis 2017.

4.2. Experimental Details

Our network is implemented in PyTorch. Adam is used as an optimizer. Transfer learning is adopted in our work. MobileNetV2 used in the encoder is pre-trained on the ImageNet, which speeds up network convergence and improves segmentation accuracy. To prevent overfitting, a strategy of changing learning rates is used in training. The learning rate is multiplied by 0.8 every 30 iterations. The initial learning rate is 3×10^{-5} on CataIS and 1×10^{-5} on EndoVis 2017. The batch sizes used on CataIS and EndoVis2017 are 8 and 16, respectively. We apply data augmentation methods such as random flip and rotation when training on EndoVis2017.

The prediction procedure of the target object is end-to-end. The input is the surgical image, and the output is the segmentation mask. At testing time, SurgiNet removes the branch of distillation and retains the backbone, attention decoder, and pyramid feature aggregation.

After experimental comparison, γ in focal loss is set to 4 for the best performance. The α in Eq. (9) is set to 1. All comparison methods are trained with the focal loss for fair comparison.

The Intersection-over-Union (IoU), standard Dice similarity coefficient (Dice), and pixel accuracy (PA) are selected as evaluation metrics. The mean values of these three metrics are represented as mIoU, mDice, and mPA.

$$Dice = \frac{2|p \cap g|}{|p| + |g|}, IoU = \frac{|p \cap g|}{|p \cup g|}, PA = \frac{|p \cap g|}{|p|} \tag{10}$$

4.3. Results on CataIS

4.3.1. Select the number of channels for the Pyramid Attention Features

The pyramid attention features are aggregated together for the final prediction, which helps adapt to scale variation. Their channel numbers will directly affect the final output result. Thus, a set of experiments is set to select the appropriate number of channels.

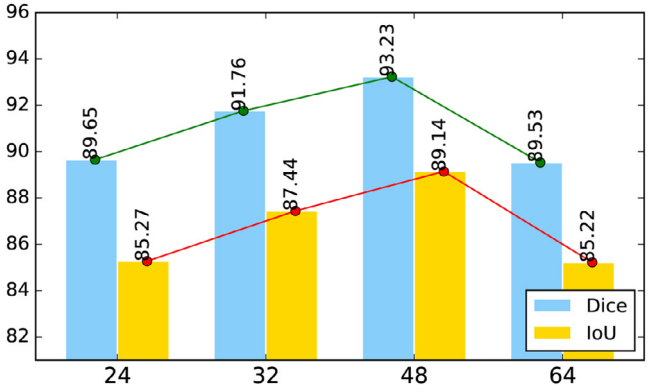

Fig. 4. Performance comparison for different channel settings of the pyramid attention features.

Table 2
Ablation study for DAM on CataIS. The baseline represents SurgiNet without DAM and CSD. It can be observed that applying DAM can significantly increase accuracy. And, the performance of DAM is better than other attention modules.

Method	Attention	mDice(%)	mIoU(%)	mPA(%)
Baseline	None	88.19	82.39	92.95
Baseline	PSAM	88.26	82.82	92.00
Baseline	Non-Local	90.97	86.00	96.00
Baseline	GCBlock	90.93	85.65	95.70
Baseline	PAB	89.73	85.46	95.89
Baseline	CAB	90.64	85.97	94.95
Baseline	DAM	91.89	87.44	96.54

Table 3
Ablation study for DAM and CSD on CataIS. The baseline represents SurgiNet without DAM and CSD.

Method	DAM	CSD	mDice(%)	mIoU(%)	mPA(%)
Baseline	×	×	88.19	82.39	92.95
Baseline	×	✓	92.10	87.49	95.07
Baseline	✓	×	91.89	87.44	96.54
Baseline	✓	✓	93.23	89.14	97.13

We test the settings of 24, 32, 48, and 64 channels, respectively. The experimental results are displayed in Fig. 4. It can be found that the best results are obtained when the number of channels is 48. Therefore, in the following experiments, the channel numbers of pyramid attention features are set to 48.

4.3.2. Ablation Study for Double Attention Module

Double attention module (DAM) models semantic dependencies between locations and channels to infer semantic features in disturbed areas, addressing illumination issues. To evaluate its performance, a series of experiments are set up. The experimental results are shown in Table 2 and Table 3.

In Table 2, the baseline is SurgiNet without DAM and CSD, which achieves 88.19% mDice and 82.39% mIoU. The baseline with DAM gets 91.89% mDice and 87.44% mIoU. By applying the DAM, the mDice increases by 3.70% and the mIoU increases by 5.05%. In Table 3, the baseline with CSD obtains 92.10% mDice and 87.49% mIoU. Compared with it, the mDice and mIoU of the baseline with both DAM and CSD have increased by 1.13% and 1.65%, respectively. Moreover, pairwise self-attention module (PSAM) (Wang et al., 2020a), non-local (Wang et al., 2018) and the GC block (Cao et al., 2019) are tested. The performance of DAM exceeds that of non-local by 0.92% mDice and 1.44% mIoU and exceeds that of GC block by 0.96% mDice and 1.79% mIoU. The accuracy of our DAM is significantly higher than PSAM, surpassing it by 3.63% mDice and 4.62% mIoU. Based on the above results, it can be found that the DAM can significantly improve segmentation accuracy.

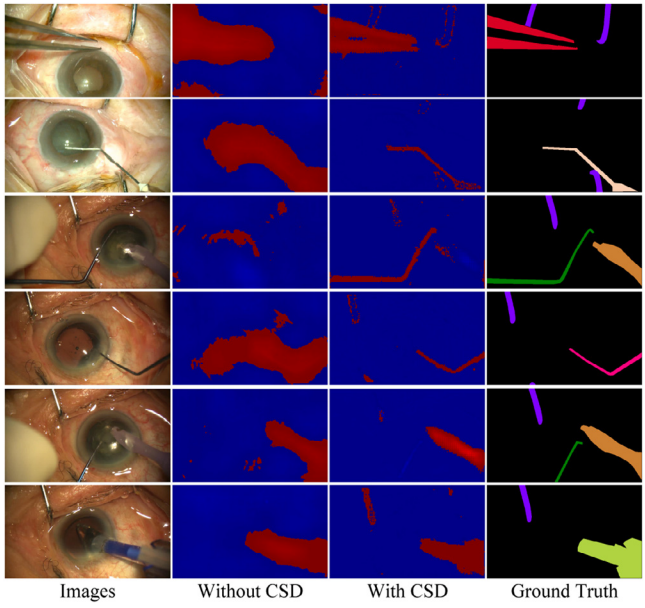

Fig. 5. Visualization of class probability map. The original images, class probability map without CSD and with CSD, as well as ground truth are shown. When applying CSD, the probability map more focuses on the surgical instrument and is close to the ground truth.

To verify the rationality of double attention, the position attention block (PAB) and the channel attention block (CAB) are tested separately. As shown in Table 2, the network only using PAB achieves 89.73% mDice and 85.46% mIoU. Compared with the baseline, the mDice and mIoU increase by 1.54% and 3.07%, respectively. Besides, the network only using CAB gets 90.64% mDice and 85.97% mIoU, which outperforms the baseline by 2.45% mDice and 3.58% mIoU. The above results indicate that both PAB and CAB contribute to improving the segmentation performance. Besides, the performance of the DAM is better than that of PAB and CAB. These results prove that double attention is better than a single attention block.

4.3.3. Ablation Study for Class-wise Self-Distillation

The class-wise self-distillation module (CSD) is introduced to enhance representation learning by distilling knowledge from multi-level class probability maps. Experiments are performed to confirm its effectiveness, which is shown in Table 3.

The baseline is the SurgiNet without CSD and DAM. When not using DAM, the network with CSD achieves 92.10% mDice and 87.49% mIoU. Compared with the baseline, the mDice and mIoU increase by 3.91% and 5.10%, respectively. The baseline with DAM gets 91.89% mDice and 87.44% mIoU. Applying CSD achieves 1.34% mDice and 1.70% mIoU gain. These results show that CSD can significantly improve feature representation.

To further analyze the effect of class-wise self-distillation, the probability maps of the network with CSD and without CSD are visualized, which is shown in Fig. 5. When CSD is not applied, the probability map is inaccurate, and it focuses on many irrelevant areas. When applying CSD, the probability map focuses on the surgical instrument and is close to the ground truth. The comparison shows that class-wise self-distillation can effectively enhance the representation learning of the network and calibrate the probability distribution.

4.3.4. Comparison with State-of-the-arts

SurgiNet and some state-of-the-arts methods are evaluated on CataIS. The results are shown in Table 4. It can be found that SurgiNet achieves the best performance and get 93.23% mDice and 89.14% mIoU. BARNet (Ni et al., 2020b) takes the second place and gets 91.46% mDice and 87.36% mIoU. SurgiNet outperforms BARNet by 1.77% mDice and 1.78% mIoU. DeepLabV3+ (Chen et al., 2018a) gets 89.69% mDice and 85.48% mIoU, which is inferior to SurginNet by 3.54% mDice and 3.66% mIoU. Besides, to prove the effectiveness of our improvement, PAANet (Ni et al., 2020a) is tested, which gets 88.42% mDice and 83.21% mIoU. Compared with PAANet, the performance of SurgiNet has been significantly improved. The mDice and mIoU are increased by 4.81% and 5.93%, respectively. SurgiNet also surpasses DANet (Fu et al., 2019) by 3.24% mDice and 3.68% mIoU, and exceeds PAN (Li et al., 2018) by 3.99% mDice and 4.28% mIoU. The gap between other methods and SurgiNet is more significant. The above experimental results suggest that SurgiNet achieves state-of-the-art performance on CataIS.

Flops is calculated to evaluate the computational cost of models. The parameter quantity reflects the size of the model. The proposed SurgiNet only has 2.05 M parameters and its Flops is 1.66 G. The model size of PAANet (Ni et al., 2020a) is 22.26 M, which is 10.86 times that of SurgiNet. The Flops of PAANet is 26.61 G, which is 16.03 times that of SurgiNet. Besides, DeepLabV3+ has 22.44 M parameters and its Flops is 29.70 G, which are 10.95 times and 17.89 times that of SurgiNet, respectively. The above results show that the SurgiNet can achieve advanced performance with very little computational cost. These also prove that class-wise self-distillation can enhance the representation learning of the network and significantly improve segmentation performance.

To further evaluate the segmentation performance of the proposed method for each type of surgical instrument, the mIoU and mDice for each category are calculated, which are shown in Table 5. It can be found that SurgiNet achieves excellent performance in every type of instrument. It takes first place in five categories (I1, I2, I8, I9, I10) and takes second place in two categories (I3, I7). Among all categories, primary incision knife (I1) is the most difficult category to be segmented. The primary incision knife is used for a short time in surgery. Thus, it has few samples, which leads to the under-fitting of the network. Since class-wise self-distillation can effectively enhance the representation learning of the network, the results of the SurgiNet in I1 are significantly better than other networks. Besides, some parts of the implant injector (I8) are transparent. So it is greatly affected by illumination issues. Most networks achieve poor segmentation performance in this category. Our method can capture long-range semantic dependencies to infer semantic features in disturbed areas. Thus, our method outperforms other methods by a significant margin in the implant injector (I8). These results show that SurgiNet can effectively address illumination issues and class-wise self-distillation can significantly enhance the network performance.

To give an intuitive display, the visualization results of SurgiNet and other methods are shown in Fig. 6. In images 2 and 4, various degrees of reflection are shown, which changes the surgical instruments to silvery white. In images 3, 4, 5, and 6, the illumination condition in the image is significantly darker, which makes it difficult to distinguish surgical instruments from the background. Nevertheless, SurgiNet still segments the surgical instruments very well. The masks of SurgiNet are more complete and closer to the ground truth.

4.3.5. Performance Comparison under Different Illumination Conditions

To verify the adaptability of SurgiNet to different illumination conditions, we evaluated it under different illumination conditions based on CataIS. Video 7 in the CataIS test set is the bright scene, and videos 8 and 9 are the dark scene. The experimental results are shown in Table 6.

It can be found that DeepLabV3+ (Chen et al., 2018a), RefineNet (Lin et al., 2017a) and UNet (Ronneberger et al., 2015)

Table 4
Performance comparison of various methods on CataIS. The proposed SurgiNet achieves the best performance 89.69% mDice and 89.14% mIoU with only 1.66 GFlops and 2.05 M parameters.

Method	mDice(%)	mIoU(%)	mPA(%)	Flops(G)	Parameter(M)
U-Net Ronneberger et al. (2015)	67.93	56.00	70.93	49.97	7.85
RefineNet Lin et al. (2017a)	80.03	71.68	79.69	215.39	71.38
LinkNet Chaurasia and Culurciello (2017)	81.02	73.06	82.2	19.08	21.77
PAN Li et al. (2018)	89.24	84.86	90.42	28.02	21.48
DANet Fu et al. (2019)	89.99	85.46	94.26	28.88	22.79
BARNet Ni et al. (2020b)	91.46	87.36	95.25	25.26	21.87
DeepLabV3+ Chen et al. (2018a)	89.69	85.48	90.96	29.70	22.44
PAANet Ni et al. (2020a)	88.42	83.21	88.80	26.61	22.26
SurgiNet(Ours)	93.23	89.14	97.13	1.66	2.05

Table 5
Performance comparison of different methods in all categories on CataIS. The proposed SurgiNet takes the first place in five categories and achieves the second place in two categories.

Method	Metric	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11
U-Net	IoU(%)	4.64	67.57	57.27	45.47	76.99	24.32	79.16	80.12	40.91	49.23	90.28
	Dice(%)	8.87	80.64	72.83	62.52	87.00	39.12	88.37	88.96	58.07	65.98	94.89
RefineNet	IoU(%)	13.00	85.67	71.17	81.58	80.94	29.31	97.65	61.62	91.41	77.67	98.51
	Dice(%)	23.01	92.28	83.16	89.85	89.46	45.34	98.81	76.25	95.51	87.43	99.25
LinkNet	IoU(%)	5.96	78.56	79.51	70.80	92.04	51.71	90.07	63.00	96.22	76.17	99.60
	Dice(%)	11.25	87.99	88.58	82.90	95.85	68.17	94.77	77.30	98.07	86.47	99.80
BARNet	IoU(%)	26.28	83.99	**96.40**	**93.89**	**99.32**	73.79	**99.38**	96.10	97.06	95.39	99.35
	Dice(%)	41.62	91.30	**98.17**	**96.85**	**99.66**	84.92	**99.69**	98.01	98.51	97.64	99.67
DeepLabV3+	IoU(%)	15.26	80.40	92.42	93.64	98.22	**85.66**	98.81	93.29	94.84	89.01	98.69
	Dice(%)	26.49	89.14	96.06	96.71	99.10	**92.27**	99.40	96.53	97.35	94.19	99.34
PAANet	IoU(%)	15.92	85.15	92.08	86.70	96.90	72.41	97.63	81.94	92.15	94.85	**99.61**
	Dice(%)	27.47	91.98	95.87	92.88	98.43	84.00	98.80	90.07	95.92	97.36	**99.80**
SurgiNet	IoU(%)	**55.87**	**98.36**	93.20	93.53	93.13	50.52	99.13	**98.68**	**99.59**	**99.08**	99.44
	Dice(%)	**71.69**	**99.17**	96.48	96.66	96.44	67.12	99.57	**99.33**	**99.79**	**99.54**	99.72

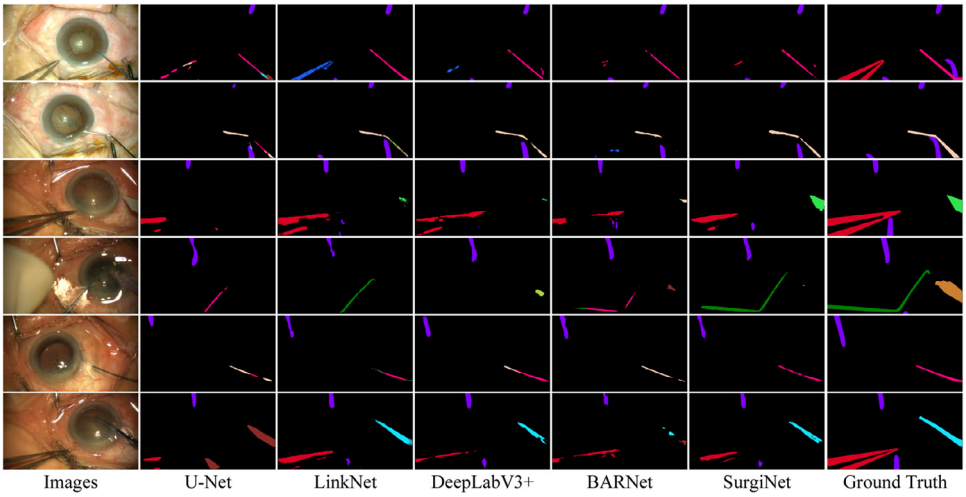

Fig. 6. Segmentation results of various methods on CataIS. It can be found that the prediction of SurgiNet is more complete and has fewer recognition errors than that of other methods.

Table 6
Performance comparison under different illumination conditions based on CataIS.

Method	Bright Scene		Dark Scene	
	mIoU(%)	mDice(%)	mIoU(%)	mDice(%)
Unet	65.80	75.76	35.84	42.55
LinkNet	72.18	79.38	67.40	75.42
RefineNet	75.92	82.87	35.84	42.55
DeepLabV3+	89.40	93.12	80.05	84.83
BARNet	90.28	94.43	84.04	87.58
SurgiNet(Ours)	92.02	95.58	84.50	89.21

have poor performance in dark scenes. Different from them, our SurgiNet achieves excellent results in both bright and dark scenes. Compared with DeepLabV3+, the performance of SurgiNet increases by 2.46% mDice and 2.62% mIoU in bright scenes and increases by 4.38% mDice and 4.45% mIoU in dark scenes. Besides, the SurgiNet also surpasses BARNet by 1.15% mDice and 1.74% mIoU in bright scenes, and it also exceeds BARNet by 1.63% mDice and 0.46% mIoU in dark scenes.

4.4. Results on EndoVis 2017

To further verify the performance of the proposed network, it is also evaluated on the EndoVis 2017 dataset (Allan et al., 2019). The experimental results on the test set of EndoVis 2017 are shown in Table 7. The proposed SurgiNet achieves 66.30% mIoU, outperforming existing methods by a large margin. PAANet (Ni et al., 2020a) only achieves 64.20% mIoU, which is inferior to SurgiNet by 2.10% mIoU. Compared with DeepLabV3+ (Chen et al., 2018a),

Table 7
Performance comparison of SurgiNet and various methods on EndoVis 2017.

Method	BiForcep	ProForcep	Driver	Sealer	Retractor	Scissor	mIoU	Flops(G)	Para.(M)
Unet	68.6	52.3	84.4	19.3	0.0	51.8	46.1	66.63	7.85
TernausNet	66.4	65.0	91.6	42.5	0.0	72.8	56.4	137.42	25.36
LinkNet	77.5	55.0	93.3	42.4	0.0	67.9	56.0	25.44	21.77
PAN	78.7	65.9	97.7	65.3	0.0	77.1	64.1	37.36	21.48
PAANet	77.5	69.2	**98.5**	64.9	0.0	75.0	64.2	35.47	22.26
DANet	74.4	63.8	96.6	66.3	0.0	77.3	63.1	38.51	22.79
DeepLabV3+	74.6	70.9	97.3	65.7	0.0	76.7	64.2	7.65	4.38
SurgiNet(Ours)	**80.4**	**71.6**	98.2	**70.4**	0.0	**77.4**	**66.3**	2.21	2.05

Table 8
Ablation study for DAM and CSD on EndoVis 2017. The baseline represents SurgiNet without DAM and CSD.

Method	DAM	CSD	mIoU(%)
Baseline	×	×	56.6
Baseline	×	✓	59.1
Baseline	✓	×	63.4
Baseline	✓	✓	66.3

the mIoU of our network has increased by 2.10%, which is a significant margin. Besides, the winner of the challenge that year, TernausNet (Iglovikov and Shvets, 2018), achieves 56.40% mIoU. SurgiNet outperforms it by 9.90% mIoU. The performance of other methods is much poorer than our SurgiNet. Besides, to evaluate the performance of SurgiNet in different categories, the mIoU in each category is also shown in Table 7. It can be found that our SurgiNet achieves first place in four categories. In other categories, our SurgiNet also achieves relatively high accuracy. Based on the above results, we can prove that our network gets state-of-the-art performance on this dataset.

To evaluate the computational cost of models, we calculate the Flops of different models. The amount of parameters is also calculated, which reflects the size of the model. The proposed SurgiNet only has 2.05 M parameters and its Flops is 2.21 G. The Flops of PAANet is 35.47 G, far exceeding that of SurgiNet. Besides, the TernausNet has 25.36 M parameters and its Flops is 137.42 G, which are 12.37 times and 62.18 times that of SurgiNet, respectively. The above results show that the computational cost of our SurgiNet is very small.

To further verify the effectiveness of DAM and CSD on EndoVis2017, a series of ablation experiments are performed and the results are displayed in Table 8. The baseline is the SurgiNet without CSD and DAM. The baseline achieves 56.60% mIoU. By applying DAM, the mIoU increased by 6.80%. Applying CSD achieves a 2.50% mIoU gain. When using both DAM and CSD, the mIoU increased by 9.70%. These results prove the effectiveness of DAM and CSD on EndoVis 2017.

To give a more intuitive display, the segmentation results of SurgiNet are visualized in Fig. 7. There are serious specular reflections in images 1 and 5. And there are many shadows in images 2, 3, and 4. Nevertheless, the SurgiNet can still segment all surgical instruments well. The segmentation result of our network is very close to the ground truth. The shape details in the tip of the surgical instrument are also very accurate. This fully proves that the double attention module can effectively solve the illumination issue.

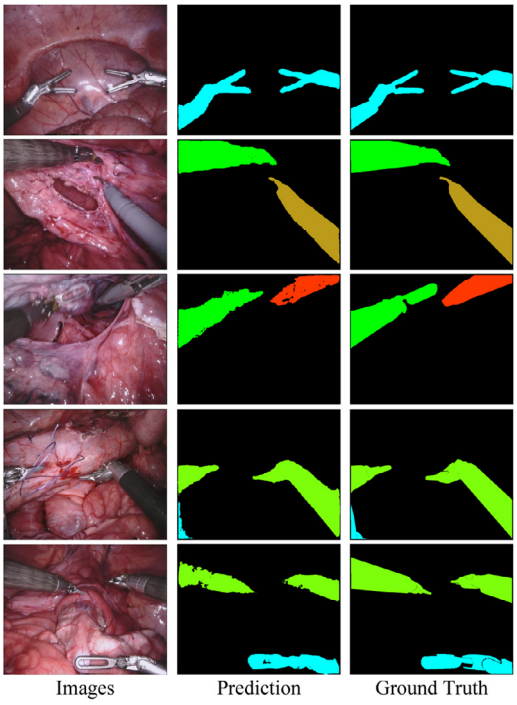

Fig. 7. Visualization for segmentation results of SurgiNet on EndoVis 2017.

5. Discussion

Based on the above quantitative and qualitative results, it can be found that SurgiNet achieves state-of-the-art performance with a small computational cost and model size. This is because class-wise self-distillation performs knowledge distillation based on probability maps among multiple layers, which enhances the representation learning of the network. Because the learning ability of the network is enhanced, we can use a more lightweight backbone to achieve higher performance, which facilitates its deployment on surgical robots. The class-wise self-distillation is also easy to apply to other networks for model compression.

The double attention module captures semantic dependencies between both locations and channels. It can infer the semantic features in the disturbed area based on semantic dependence. Experiments show that it can significantly improve the performance of the network. Moreover, as shown in Table 2, double attention is better than any single attention. These results prove the rationality of the double attention module. Besides, surgical instruments are constantly moving during the surgery, their size and shape will be changed drastically. The aggregation of pyramid attention features contributes to the segmentation of objects with different scales. The network can capture local details from large-scale feature maps while capturing overall shape from small-scale feature maps, improving the feature representation. In this way, the shape and size features of surgical instruments at different scales can be learned to recognize them correctly. Note that the segmentation performance is sensitive to the number of channels for the attention features, which is demonstrated in Fig. 4. Lightweight network, MobileNetV2 (Sandler et al., 2018), is adopted as the backbone of our SurgiNet, and the number of channels is set to 48 for

the best performance. When the backbone is replaced, the channel number of the attention feature should also be re-selected to adapt to the backbone.

6. Conclusion

In this paper, the SurgiNet is proposed to learn pyramid attention and distill knowledge from itself for surgical instrument segmentation. The double attention module is designed to capture semantic dependencies between locations and channels. It can infer semantic features in the areas affected by illumination variation, addressing the illumination issue. Pyramid attention features are aggregated for final prediction, which helps to adapt to scale variation. Besides, class-wise self-distillation makes the network extract knowledge from itself based on class probability map, enhancing the representation learning ability of the network. Experiments have proved their effectiveness. The proposed network achieves state-of-the-art performance on both CataIS and EndoVis 2017.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant 62027813, U20A20196, U1713220, U1913601, 62003343), the National Key Research and Development Program of China (Grant 2020YFF01014800ZL), the CAS Interdisciplinary Innovation Team (JCTD-2019-07), the Youth Innovation Promotion Association of the Chinese Academy of Sciences (Grant 2018165).

References

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., Garcia-Peraza-Herrera, L., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M., 2019. 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426.

Attia, M., Hossny, M., Nahavandi, S., Asadi, H., 2017. Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 3373–3378.

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. arXiv preprint arXiv:1904.11492.

Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, pp. 1–4.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J., 2018. A^2-nets: Double attention networks. In: Advances in Neural Information Processing Systems 31, pp. 352–361.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.

Hou, Y., Ma, Z., Liu, C., Loy, C.C., 2019. Learning lightweight lane detection cnns by self attention distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1013–1021.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. doi:10.1109/CVPR.2018.00745.

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746.

Islam, M., Vibashan, V., Lim, C.M., Ren, H., 2021. St-mtl: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. Medical Image Analysis 67, 101837.

Jin, Y., Cheng, K., Dou, Q., Heng, P.-A., 2019. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 440–448.

Jin, Y., Cheng, K., Dou, Q., Heng, P.-A., 2019. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: Medical Image Computing and Computer Assisted Intervention, pp. 440–448.

Kim, J.-H., Jun, J., Zhang, B.-T., 2018. Bilinear attention networks. In: Advances in Neural Information Processing Systems, pp. 1564–1574.

Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W., Zhang, B.-T., 2016. Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325.

Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.

Li, J., Li, J., Fang, F., Li, F., Zhang, G., 2020. Luminance-aware pyramid network for low-light image enhancement. IEEE Transactions on Multimedia PP (99). 1–1

Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5168–5177.

Lin, T., Goyal, P., Girshick, R., He, K., Dollr, P., 2017. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007.

Ni, Z.-L., Bian, G.-B., Wang, G., Zhou, X.-H., Hou, Z.-G., Chen, H.-B., Xie, X.-L., 2020. Pyramid attention aggregation network for semantic segmentation of surgical instruments. In: AAAI, pp. 11782–11790.

Ni, Z.-L., Bian, G.-B., Wang, G.-A., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Li, Z., Wang, Y.-H., 2020. Barnet: Bilinear attention network with adaptive receptive fields for surgical instrument segmentation. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pp. 832–838.

Ni, Z.-L., Bian, G.-B., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Wang, C., Zhou, Y.-J., Li, R.-Q., Li, Z., 2019. Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In: International Conference on Neural Information Processing. Springer, pp. 139–149.

Qin, F., Li, Y., Su, Y., Xu, D., Hannaford, B., 2019. Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 9821–9827.

Qin, F., Lin, S., Li, Y., Bly, R.A., Moe, K.S., Hannaford, B., 2020. Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision. arXiv preprint arXiv:2002.10675.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520.

Sarikaya, D., Corso, J.J., Guru, K.A., 2017. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE Transactions on Medical Imaging 36 (7), 1542–1549. doi:10.1109/TMI.2017.2665671.

Wang, C., Wu, Y., Su, Z., Chen, J., 2020. Joint self-attention and scale-aggregation for self-calibrated deraining network.

Wang, G., Gong, S., Cheng, J., Hou, Z., et al., 2020. Faster person re-identification. In: European Conference on Computer Vision. Springer, pp. 275–292.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794–7803. doi:10.1109/CVPR.2018.00813.

Wang, Y., Zhou, W., Jiang, T., Bai, X., Xu, Y., 2020. Intra-class Feature Variation Distillation for Semantic Segmentation. Springer, Cham.

Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: The IEEE International Conference on Computer Vision (ICCV), pp. 1821–1830.

Yun, S., Park, J., Lee, K., Shin, J., 2020. Regularizing class-wise predictions via self-knowledge distillation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K., 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3713–3722.

Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G., 2018. Progressive attention guided recurrent network for salient object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 714–722. doi:10.1109/CVPR.2018.00081.

Zhang, Z., Sabuncu, M. R., 2020. Self-distillation as instance-specific label smoothing.

Zhou, C., Chen, Y., Fan, M., Wen, Y., Chen, H., Chen, L., 2020. Enhancing tiny tissues segmentation via self-distillation. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 934–940.