Pyramid Attention Aggregation Network for Semantic Segmentation of Surgical Instruments

Zhen-Liang Ni,[1,2] Gui-Bin Bian,[1,2*] Guan-An Wang,[1,2] Xiao-Hu Zhou,[2]
Zeng-Guang Hou,[1,2,3] Hua-Bin Chen,[1,2] Xiao-Liang Xie[2]
[1]University of Chinese Academy of Sciences, Beijing 100049, China
[2]State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China
{nizhenliang2017, guibin.bian, wangguanan2015, xiaohu.zhou, zengguang.hou, chenhuabin2019, xiaoliang.xie}@ia.ac.cn

Abstract

Semantic segmentation of surgical instruments plays a critical role in computer-assisted surgery. However, specular reflection and scale variation of instruments are likely to occur in the surgical environment, undesirably altering visual features of instruments, such as color and shape. These issues make semantic segmentation of surgical instruments more challenging. In this paper, a novel network, Pyramid Attention Aggregation Network, is proposed to aggregate multi-scale attentive features for surgical instruments. It contains two critical modules: Double Attention Module and Pyramid Upsampling Module. Specifically, the Double Attention Module includes two attention blocks (i.e., position attention block and channel attention block), which model semantic dependencies between positions and channels by capturing joint semantic information and global contexts, respectively. The attentive features generated by the Double Attention Module can distinguish target regions, contributing to solving the specular reflection issue. Moreover, the Pyramid Upsampling Module extracts local details and global contexts by aggregating multi-scale attentive features. It learns the shape and size features of surgical instruments in different receptive fields and thus addresses the scale variation issue. The proposed network achieves state-of-the-art performance on various datasets. It achieves a new record of 97.10% mean IOU on Cata7. Besides, it comes first in the MICCAI EndoVis Challenge 2017 with 9.90% increase on mean IOU.

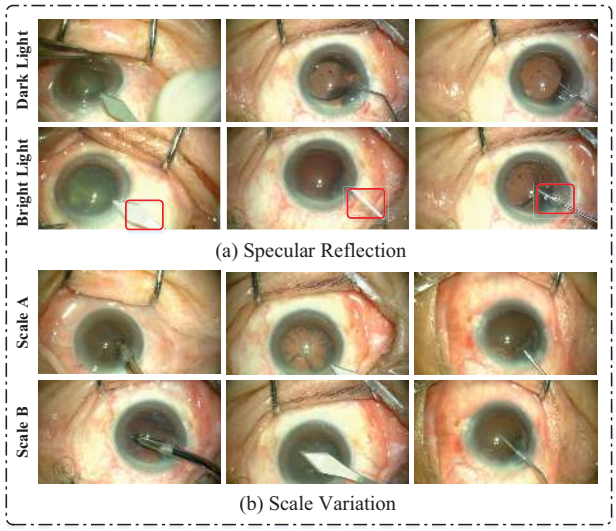

Figure 1: Difficulties in semantic segmentation for surgical instruments. (a) Specular Reflection: each column is the same instrument but under different illustrations. As we can see, under highlights, its appearance is white otherwise dark. (b) Scale Variation: shape and size of the same instrument are different due to poses and views.

Introduction

In recent years, significant progress has been witnessed 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, Corso, and Guru 2017). By segmenting the surgical instruments and estimating their poses, the navigation and control for surgical robots can be assisted. Also, this technology can give real-time warnings during surgery and reduce the risk of surgery. 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, Corso, and Guru 2017). These applications can improve the safety of surgery and reduce the workload of doctors, which is significant for clinical work.

However, semantic segmentation of surgical instruments is very different from common segmentation tasks and faces more difficulties. On the one hand, as shown in Figure 1(a), surgery usually requires intense lighting conditions, which leads to specular reflection and affects the appearance of the surgical instruments. For example, the surgical instruments are more white when illuminated. On the other hand, since the surgical instrument is constantly moving during the surgery, its size and shape are always changing, which is shown in Figure 1(b). These two problems make the seman-

*Corresponding Author

tic segmentation of surgical instruments more challenging.

Recently, many methods have been proposed for semantic segmentation of surgical instruments. A hybrid CNN-RNN method (Attia et al. 2017) introduced Recurrent Neural Network to capture the global context and expand the receptive field, improving the feature representation. MF-TAPNet (Jin et al. 2019) utilized the inherent temporal clues from the instrument motion to improve segmentation results. It inferred a prior indicating the instrument's location and shape via optical flow. Another work (Qin et al. 2019) fused the prediction of Convolutional Neural Network and the kinematic pose information to boost segmentation accuracy. However, those works mainly focus on fusing different forms of information for higher segmentation accuracy while fails to explicitly deal with the specular reflection and scale variation, limiting their performances.

To address the issues mentioned above, we rethink the changes in visual features brought by these two challenges. Specular reflection affects visual features such as color and texture of surgical instruments. Thus, it is difficult for the network to identify surgical instruments based on these features directly. The network needs to infer the semantic features of these regions from neighboring pixels based on global contexts and semantic dependencies. Besides, the scale variation of surgical instruments affects visual features such as shape and size. Aggregation for multi-scale features is conducive to capturing the shape and size features of surgical instruments at different scales. To this end, the Pyramid Attention Aggregation Network (PAANet) is proposed to learn discriminative features for surgical instruments. The proposed PAANet contains the Double Attention Module (DAM) to model semantic dependencies between positions and channels and the Pyramid Upsampling Module (PUM) to aggregate multi-scale attentive features.

Specifically, the Double Attention Module consists of position attention block and channel attention block. The position attention block captures joint semantic information to model semantic dependencies between positions. The channel attention block squeezes global information into a channel-attention vector, which encodes the semantic dependencies between channels. Their outputs are fused and calibrated to generate attentive features. The attentive features can boost the distinction between semantic features and emphasize the target regions, contributing to addressing specular reflection. Besides, the Double Attention Module takes very little computational cost. Thus, it can be easily inserted into other models to improve their performance.

The Pyramid Upsampling Module aggregates multi-scale attentive features by performing pyramid upsampling and concatenation. It captures local details from large-scale feature maps while captures global contexts from small-scale feature maps, improving the feature representation. In this way, the network can learn the shape and size features of surgical instruments at different scales, helping to address the scale variation issue. Besides, attentive features contain rich semantic information. Due to the upsampling and skip connection, the semantic information contained in attentive feature maps at different scales is distinguishing. The aggregation of multi-scale attentive features can integrate multi-scale semantic information to supplement the information lost during the delivery process, making the final predictions more reliable.

The contributions of this work can be concluded as follows:

- The Double Attention Module captures semantic dependencies between positions and channels to generate attentive features. The attentive features can distinguish target regions, contributing to solving the specular reflection issue.
- The Pyramid Upsampling Module captures local details and global contexts by aggregating multi-scale attentive features. Also, it integrates multi-scale semantic information to make predictions more reliable.
- The proposed network achieves state-of-the-art performance on various datasets. It achieves a new record of 97.10% mean IOU on Cata7 and comes first in the MICCAI EndoVis Challenge 2017 with 9.90% increase on mean IOU.

Related Work

Semantic Segmentation of Surgical Instrument

Recently, Fully Convolutional Network (FCN) is widely used in semantic segmentation of surgical instruments due to their excellent performance on feature extraction. Some work modified the architecture of the FCN to improve accuracy (Laina et al. 2017; Shvets et al. 2018). For example, ToolNet (García-Peraza-Herrera et al. 2017) modified FCN-8s by adding feature maps at different scales in a cascaded fashion to acquire refined edge of surgical instruments. Another way is to introduce other methods such as optical flow and recurrent neural network (RNN) to improve the performance of FCN. For example, a work (García-Peraza-Herrera et al. 2016) applied optical flow to reduce computational complexity and improve the performance of the network. The hybrid CNN-RNN method (Attia et al. 2017) introduced Recurrent Neural Network to capture the global context and expand the receptive field, improving the feature representation. Drawing on this idea, the attention mechanism is introduced to capture global contexts and semantic dependencies, addressing the specular reflection issue.

Attention Used in Semantic Segmentation

The attention model focuses on key regions by mimicking human attention mechanisms. Recently, it has been widely used in semantic segmentation tasks. Attention mechanisms include the position attention mechanism (Wang et al. 2018; Chen et al. 2018b; Cao et al. 2019) and the channel attention mechanism (Hu, Shen, and Sun 2018; Li et al. 2018), which model semantic dependencies between positions and channels, respectively. Non-local block (Wang et al. 2018) captured long-range dependencies in space without being limited by distance, which contributes to scene understanding. The Squeeze-and-Excitation block (Hu, Shen, and Sun 2018) captured global information to model semantic dependencies between channels. To capture semantic information

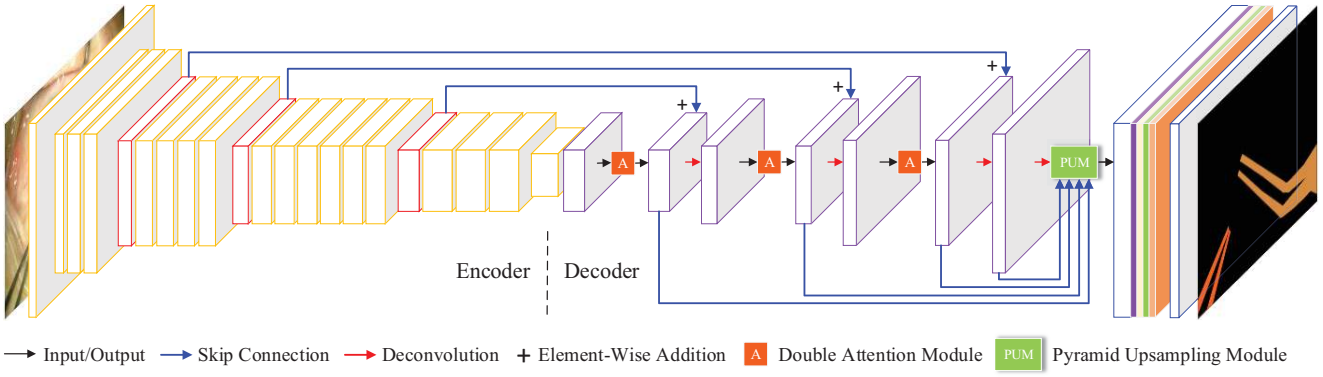

Figure 2: The architecture of Pyramid Attention Aggregation Network. It adopts an encoder-decoder architecture. ResNet-34 pre-trained on the ImageNet is adopted as an encoder. The decoder consists of Double Attention Module, Pyramid Upsampling Module, and Deconvolution. The output of the PAANet is the same size as the original image.

between channels and positions simultaneously, some models fused these two attention mechanisms, such as Dual Attention Network (Fu et al. 2018) and Progressive Attention Guidance Module (Zhang et al. 2018). These models show that the attention mechanism contributes to improving the feature representation.

Aggregation of Multi-Scale Features

Several approaches aggregate multi-scale features to capture different scale objects. PSPNet (Zhao et al. 2017) used spatial pyramid pooling to aggregate multi-scale features. Atrous spatial pyramid pooling(ASPP) utilized dilated convolution with different rates to generate feature maps with different receptive fields (Chen et al. 2018a; Chen et al. 2017). DenseASPP (Yang et al. 2018) introduced dense connections in ASPP to cover a larger scale range and improve information flow. Based on these efforts, we can see that the aggregation of multi-scale features contributes to addressing the scale variation issue.

Pyramid Attention Aggregation Network

In this section, the architecture of the Pyramid Attention Aggregation Network is first illustrated in Figure 2. Then, the principle of the Double Attention Module is illustrated in detail. Finally, how to aggregate multi-scale attentive features by Pyramid Upsampling Module is described.

Double Attention Module

Double Attention Module contains two attention mechanisms to model semantic dependencies between positions and channels, whose architecture is depicted in Figure 3. It consists of position attention block and channel attention block, which models semantic dependencies between positions and channels, respectively. Their outputs are fused and calibrated to improve the feature representation.

Position Attention Block In semantic segmentation task, we hope that the network only focuses on the target region. However, due to the specular reflection issue, it is difficult for the network to locate surgical instruments. To address this problem, the position attention block models semantic dependencies between positions to boost the distinction of semantic features and emphasize the target region. It is based on a variant of low-rank bilinear pooling to aggregate joint semantic information into a position-attention map with only one channel. The position-attention map encodes the semantic dependencies between positions to boost the distinction of semantic features.

Bilinear models are often used to capture attentive features (Chen et al. 2018b; Kim, Jun, and Zhang 2018). These previous work (Kim et al. 2016; Yu et al. 2017) propose low-rank bilinear pooling to approximate bilinear pooling by matrix decomposition and Hadamard product, which significantly reduces computational cost. Also, it can learn joint feature representation, modeling complex semantic dependencies. It is illustrated in Eq.(1).

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

where $\otimes$ refers to Hadamard product. $x \in R^N$ and $y \in R^M$ represent the input vector. $z \in R^C$ represents the output vector. $U \in R^{N \times k}$ and $V \in R^{M \times k}$ are linear projections. $P \in R^{k \times C}$ is used to control the length of the output. $b \in R^C$ is the bias vector.

In this paper, a variant of low-rank bilinear pooling is developed to generate a position-attention map, which is described in Eq.(2). 1×1 convolutions are used for linear projection to replace matrices U and V. Non-linear activations contribute to improving the feature representation of the network (Kim et al. 2016). Therefore, ReLU is used to add non-linear activations. Softmax is adopted to normalize the feature map. Sum pooling is adopted to adjust the dimension of output, which corresponds to the matrix P in Eq.(1). Finally, the position-attention map is transformed by 1×1 convolution.

The architecture of position attention block is shown in Figure 3. To reduce computational cost and aggregate information across channels, the dimension of the input feature map is reduced to C/r by 1×1 convolution, where C is the dimension of input. r can be selected according to the dimension of input. It is set to 2 in this work. $x, y \in R^{C \times W \times H}$

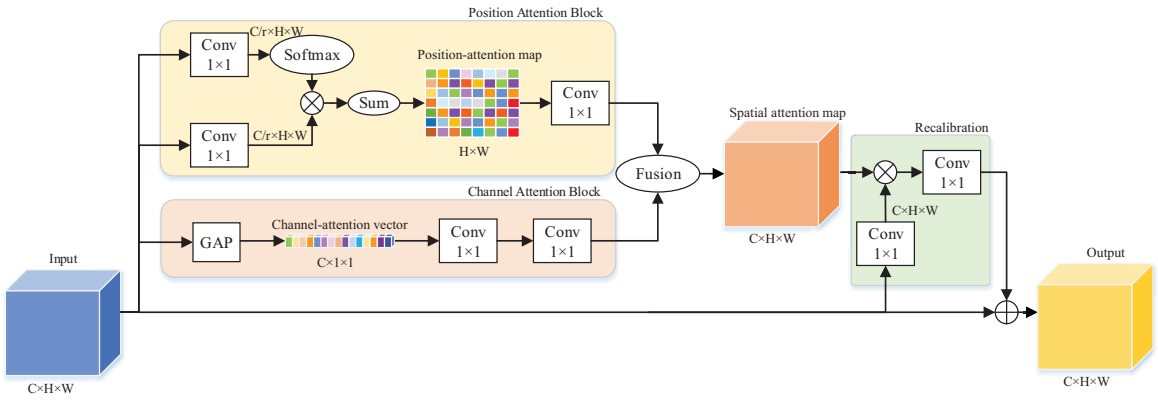

Figure 3: The architecture of the Double Attention Module. It consists of two attention blocks: position attention block and channel attention block. The outputs of these two blocks are fused to generate the spatial attention map. Recalibration block is applied to calibrate the spatial attention map and further extract semantic dependencies. $\otimes$ refers to the Hadamard product.

are the input feature maps, where H and W are the height of and the width of the feature map respectively. In this paper, we set $x = y$. $A_p \in R^{W \times H}$ is the position-attention map. U and V represent 1×1 convolution.

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

where f denotes softmax function and g denotes sum pooling. δ refers to ReLU function. Softmax is illustrated in following:

$$S_k^{w,h} = \frac{e^{\alpha_k^{w,h}}}{\sum_{k=1}^{C/r} e^{\alpha_k^{w,h}}} \tag{3}$$

where α refers to the input feature map. S represents the output of softmax. $w = 1, 2, ..., W$ and $h = 1, 2, ..., H$. Sum pooling is described as follows:

$$P_{w,h} = \sum_{k=1}^{C/r} e^{\beta_k^{w,h}} \tag{4}$$

where P refers to the output of sum pooling. $P \in R^{W \times H}$. Due to the use of sum pooling and Hadamard product, the computational cost of position attention block is very low.

Channel Attention Block Different channels correspond to the various semantic response. Surgical instruments and human tissues are often emphasized in different channels. By utilizing semantic dependencies between channels, we can emphasize specific semantic features and suppressing useless ones. Besides, since the receptive field of the convolution operation is local, the network cannot capture global semantic features. This local feature representation leads to poor semantic understanding. Thus, channel attention block is designed to model semantic dependencies between channels and capture the global context. It is based on the global average pooling.

The global average pooling aggregates global information into an attentive vector that encodes the semantic dependencies between the channels. Each element in the attention vector contains global information, contributing to expending the receptive field. The global average pooling is described in Eq.(5).

$$a_k = \varphi(x_k) = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} x_k(i, j) \tag{5}$$

where φ refers to the global average pooling. $x_k \in R^{W \times H}$ are the input feature maps. $k = 1, 2, ..., C$. $A_c = [a_1, a_2, ..., a_C]$ is the channel-attention vector. To further capture the semantic relationship between channels, the channel-attention vector is transformed by two 1×1 convolutions with batch normalization and softmax.

Fusion of Attentive Features Existing work (Fu et al. 2018; Zhang et al. 2018) performs the position attention mechanism and the channel attention mechanism separately. Different from them, we fuse position-attention map and channel-attention vector before calibration. In this way, we can effectively use attentive information and avoid introducing interference. Position-attention map and channel-attention vector are merged into a spatial attention map A_s in the following way:

$$A_s(k, w, h) = A_p(w, h) \times A_c(k) \tag{6}$$

where $k = 1, 2, ..., C$, $w = 1, 2, ..., W$ and $h = 1, 2, ..., H$.

Recalibration To further extract semantic dependencies and improve the feature representation, the spatial attention map is calibrated. The calibration consists of two steps, including semantic information calibration and nonlinear transformation. First, the semantic information in the original feature map is used to calibrate the spatial attention map, which is shown in Eq.(7).

$$\widehat{A_s} = A_s \otimes \delta(\theta x) \tag{7}$$

where θ denotes the 1×1 convolution. x is the input feature map. Then, to further capture the dependencies between spatial pixels, 1×1 convolution with ReLU is performed to transform the spatial attention map. In the experiment, we found that recalibration is crucial. It can improve the performance of the Double Attention Module.

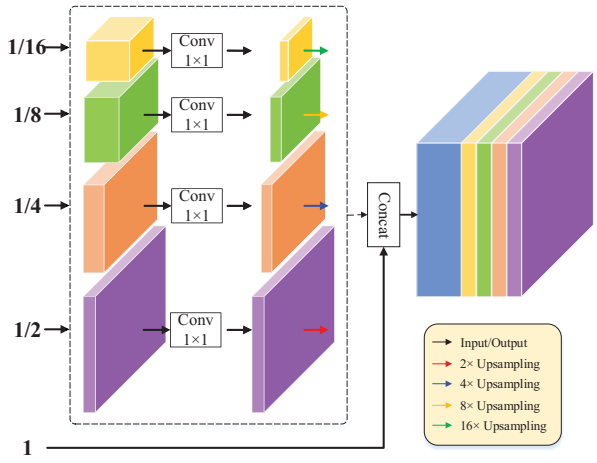

Figure 4: Pyramid Upsampling Module.

Advantages The proposed Double Attention Module is based on a variant of low-rank bilinear pooling and global average pooling to capture complex semantic dependencies. It can significantly improve segmentation accuracy while taking up very little computational cost. This is more efficient than the classic convolution operation. Moreover, it can be plugged into other convolutional networks directly to improve their performance.

Pyramid Upsampling Module

Since the surgical instrument is constantly moving during the surgery, its scale and shape are always changing, making segmentation more challenging. To address this issue, the Pyramid Upsampling Module is proposed to aggregate multi-scale attentive features. It captures local details from large-scale feature maps while captures global contexts from small-scale feature maps, improving the feature representation. Furthermore, attentive features contain rich semantic information. Due to the upsampling and skip connection, the semantic information contained in attentive feature maps at different scales is distinguishing. Aggregating multi-scale attentive features can integrate multi-scale semantic information to supplement the information lost during the delivery process, making the final predictions more reliable. The architecture of the Pyramid Upsampling Module is illustrated in Figure 4.

Specifically, the Pyramid Upsampling Module performs pyramid upsampling and concatenation to aggregates multi-scale attentive features. To reduce the computational cost and aggregate semantic information across channels, 1×1 convolution is used to reduce the dimension of feature maps to N/4. N refers to the dimension of the maximum feature map. Then, small-scale feature maps are directly upsampled to the same size as the input image via bilinear interpolation. Finally, these feature maps are concatenated as pyramid features. Due to containing rich semantic information at different scales, pyramid features are conducive to addressing the scale variation issue. The number of scales and the dimension of small-scale features can be modified. The choice of these parameters depends on the architecture of the network.

Pyramid Upsampling Module is described in Eq.(8). The input is attentive feature maps generated by Double Attention Module. x_k represents the k-th layer attentive feature map.

$$x = H([f_{2^k}(x_0), f_{2^{k-1}}(x_1), ..., f_{2^1}(x_{k-1})]) \quad (8)$$

where f_{2^k} denotes the 1×1 convolution and the $2^k\times$ upsampling, shown in Figure 4. $H(\cdot)$ refers to concatenation. In this way, the Pyramid Upsampling Module aggregates multi-scale features and integrates multi-scale semantic information.

Experiments And Results

The proposed PAANet is evaluated on the Cata7 dataset and the MICCAI EndoVis 2017 dataset. It achieves a new record of 97.10% mIOU on Cata7 and comes first in the MICCAI EndoVis Challenge 2017 with 9.90% increase on mIOU.

Dataset

Cata7 Cata7 is a cataract surgical instrument dataset for semantic segmentation. It contains 2500 frames with a resolution of 1920×1080, which contains 1800 frames for training and 700 frames for the test. These images are split from 7 cataract surgery videos. There are 10 types of surgical instruments in Cata7.

MICCAI EndoVis 2017 Dataset EndoVis 2017 is from the MICCAI Endovis Challenge 2017 (Allan et al. 2019). This dataset is acquired from a Vinci Xi robot. It contains 3000 images with a resolution of 1280×1024, including 1800 images for training and 1200 images for the test. There are 7 types of surgical instruments in EndoVis 2017.

Loss Function

To solve the class imbalance problem, we use a hybrid loss that consists of cross entropy and Jaccard (Iglovikov and Shvets 2018). As shown in Eq.(9), it merges cross entropy and Jaccard in a new way. Cross entropy is often used for semantic segmentation. However, due to the class imbalance, pixels will be misclassified into classes which contain more samples. Jaccard estimates the similarity between the prediction and the ground truth. It is not affected by the class imbalance issue. This hybrid loss retains this property of Jaccard, contributing to solving the class imbalance issue.

$$Loss = E - \alpha \log(J), \alpha \in [0, 1] \quad (9)$$

where E refers to cross entropy and J refers to Jaccard. α is a weight coefficient that balances cross entropy and Jaccard. After many experiments, α is set to 0.3 because the loss has the best performance at this time.

Experimental Details

Our network is implemented in PyTorch. Adam is used as an optimizer. The batch size is 8. To prevent overfitting, a strategy of changing learning rates is used in training. The initial learning rate is multiplied by 0.8 every 30 iterations. The initial learning rate is 6×10^{-6} on Cata7 and 3×10^{-5}

on EndoVis 2017. Due to limited computing resources, each image in Cata7 is resized to 960×544 pixels and the image in Endovis 2017 is resized to 640×512 pixels.

Transfer learning is adopted in our work. ResNet-34 used in the encoder is pre-trained on the ImageNet, which speeds up network convergence and improves segmentation accuracy. Besides, data augmentation is applied to improve sample diversity. Random rotation, shift, and flip are performed on original samples to generate new samples. The Intersection-Over-Union(IOU) and Dice are selected as the evaluation metric.

Results on Cata7

Ablation Study for Double Attention Module Double Attention Module (DAM) captures semantic dependencies between positions and channels to generate attentive features, improving the feature representation. To verify its performance, ablation experiments are performed. The experimental results are shown in Table 1 and Table 2.

In Table 1, PAANet without DAM is used as the base network. The base network achieves 98.11% mean Dice and 95.81% mean IOU. By applying DAM, the network achieves 98.82% mean Dice and 97.10% mean IOU. Mean IOU increases by 1.29% by using the DAM. In Table 2, PAANet without PUM and DAM is used as the base network, which achieves 94.63% mean Dice and 91.31% mean IOU. By employing the DAM, the mean Dice and IOU increase to 98.09% and 95.75%, respectively. These results show that the DAM can significantly improve segmentation accuracy.

Table 1: Ablation study for DAM on Cata7. DAM consists of position attention block (PAB) and channel attention block (CAB).

Method	PAB	CAB	mDice	mIOU	Param.	FLOPs
BaseNet1			98.11	95.81	21.83M	55.14G
BaseNet1		✓	98.21	95.99	22.17M	55.97G
BaseNet1	✓		98.34	96.13	22.09M	56.38G
BaseNet1	✓	✓	98.82	97.10	22.26M	56.39G

The position attention block (PAB) and the channel attention block (CAB) are evaluated separately. As shown in Table 1, applying PAB individually achieves 96.13% mean IOU, which outperforms the base network by 0.32%. Meanwhile, employing CAB individually achieves 95.99% mean IOU, which outperforms the base network by 0.18%. These results show that both PAB and CAB help to improve segmentation accuracy. Furthermore, the performance of the DAM is better than that of PAB and CAB. This indicates that the Double Attention Module can capture more discriminative attentive features than a single attention module.

The Double Attention Module has few parameters and it takes up few computational costs. Three DAMs are introduced in PAANet. As shown in Table 1, the number of parameters only increases by 0.43 M. Each DAM has only 0.143 M parameters. Furthermore, by employing the DAM, FLOPs only increases by 1.25G, which only accounts for 2.27% of the total FLOPs.

To give an intuitive display, the results are visualized in Figure 5. The red line marks the contrasted region. In the

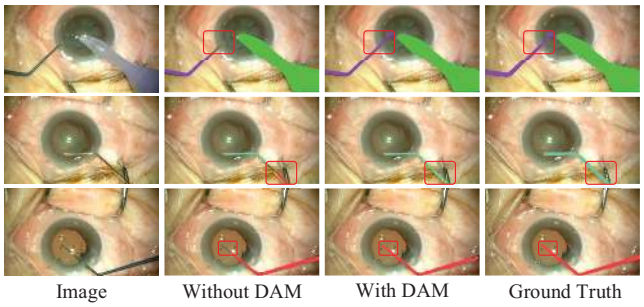

Figure 5: Visualization results of Double Attention Module (DAM) on Cata7. The red line marks the contrast region.

Table 2: Ablation study for PUM on Cata7.The BaseNet2 represents PAANet without PUM and DAM.

Method	DAM	PUM	mDice	mIOU	Param.	FLOPs
BaseNet2			94.63	91.31	21.80M	47.35G
BaseNet2		✓	98.11	95.81	21.83M	55.14G
BaseNet2	✓		98.09	95.75	22.23M	48.60G
BaseNet2	✓	✓	98.82	97.10	22.26M	56.39G

results without DAM, some surgical instruments are not entirely segmented due to the class imbalance issue. Besides, there is a misclassification in the second image, which caused by specular reflection. Meanwhile, the results with the DAM are the same as the ground truth. Since the Double Attention Module captures global contexts and semantic dependencies to boost the discrimination of semantic features, it can solve the above problems very well and help the network focus on key regions. By comparison, the effectiveness of the DAM is confirmed.

Ablation Study for Pyramid Upsampling Module The Pyramid Upsampling Module (PUM) aggregates multi-scale attentive features and integrates multi-scale semantic information. A serious of experiments are set up to confirm its effectiveness. The results are illustrated in Table 2.

PAANet without PUM and DAM is used as the base network. The network applying PUM achieves 95.81% mean IOU, which outperforms the base network by 4.50% mean IOU. When using both DAM and PUM, the network achieves 97.10% mean IOU. The mean IOU increases by 1.35% due to the use of PUM. Furthermore, PUM only adds 0.03 M parameters in PAANet, which only accounts for 0.13% of the total parameter. These results show that PUM can significantly improve segmentation accuracy without adding a mass of parameters.

Table 3: Comparison with state-of-the-art methods on Cata7.

Method	mDice	mIOU
U-Net (Ronneberger, Fischer, and Brox 2015)	86.83	78.21
RefineNet (Lin et al. 2017)	93.53	88.41
LinkNet (Chaurasia and Culurciello 2017)	94.63	91.31
TernausNet (Iglovikov and Shvets 2018)	96.40	92.98
PAANet(Ours)	98.82	97.10

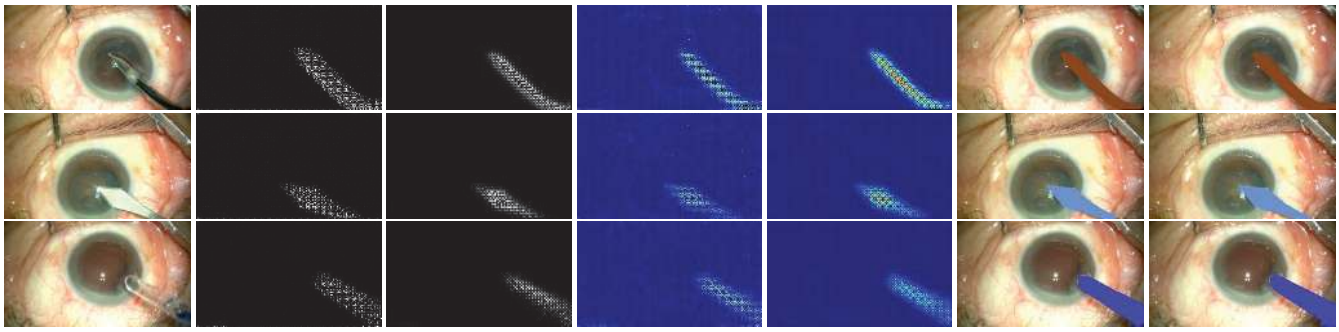

Figure 6: Visualization of attentive features and segmentation results. For each row, we first show the original image and two position-attention maps at different scales. To demonstrate the effectiveness of the channel attention block, two channels of attentive feature maps are visualized. Finally, the segmentation results and the ground truth are provided.

Table 4: Segmentation results on Endovis 2017 dataset. PAANet outperforms existing methods and achieves 64.10% mean IOU. NCT, UB, and UA are the university abbreviation of the participating team (Allan et al. 2019). Since each dataset contains a different number of samples, they are given a weight corresponding to the number of samples when calculating the mean IOU.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	mIOU
TernausNet	**0.177**	0.766	0.611	0.871	0.649	0.593	0.305	0.833	**0.357**	0.609	0.542
ToolNet	0.073	0.481	0.496	0.204	0.301	0.246	0.071	0.109	0.272	0.583	0.337
SegNet	0.138	0.013	0.537	0.223	0.017	0.462	0.102	0.028	0.315	0.791	0.371
NCT	0.056	0.499	**0.926**	0.551	0.442	0.109	0.393	0.441	0.247	0.552	0.409
UB	0.111	0.722	0.864	0.680	0.443	0.371	0.416	0.384	0.106	0.709	0.453
UA	0.068	0.244	0.765	0.677	0.001	0.400	0.000	0.357	0.040	0.715	0.346
Ours	0.106	**0.819**	0.923	**0.945**	**0.836**	**0.625**	**0.435**	**0.869**	0.318	**0.858**	**0.641**

Comparison with state-of-the-art To further verify the performance of PAANet, it is compared with state-of-the-art methods. As shown in Table 3, PAANet achieves 98.82% mean Dice and 97.10% mean IOU, outperforming other methods by a large margin. Among other methods, TernausNet has the best performance, 96.40% mean Dice and 92.98% mean IOU. PAANet exceeds TernausNet by 2.42% on mean Dice and 4.12% on mean IOU, which is a significant gap. Besides, U-Net (Ronneberger, Fischer, and Brox 2015), RefineNet (Lin et al. 2017), and LinkNet (Chaurasia and Culurciello 2017) are also evaluated on Cata7. LinkNet and RefineNet use ResNet-34 and ResNet-50 as encoders, respectively. Their performance is much poorer than that of PAANet. These results show that PAANet achieves state-of-the-art performance on Cata7.

Visualization of Attentive Feature Map Double Attention Module consists of the position attention block (PAB) and the channel attention block (CAB). PAB encodes semantic relationships into the position-attention map, which only has one channel. CAB aggregates global information into the channel-attention vector. To further verify their performance, position-attention maps and feature maps are visualized in Figure 6.

Two position-attention maps at different scales are selected for visualization, which is marked as #1 and #2. The regions of the surgical instrument are highlighted in position-attention maps, showing that PAB can effectively infer the semantic information of target regions through semantic dependencies between positions. Also, two channels of attentive feature maps are visualized to demonstrate the effect of the channel attention block. As shown in Figure 6, the regions of the surgical instrument are highlighted in the attentive feature map. These results indicate that channel attention can effectively utilize semantic dependencies between channels to emphasize specific channels, making the network focus on the target semantic information.

Results on EndoVis 2017

To verify the generalization of our proposed network, it is also evaluated on the Endovis 2017 dataset (Allan et al. 2019). The test set consists of 10 video sequences. Dataset 1-8 contain 75 images, respectively. Dataset 9 and 10 contain 300 images, respectively. Each sequence contains specific surgical instruments. The results are reported in Table 4. TernausNet (Iglovikov and Shvets 2018), ToolNet (García-Peraza-Herrera et al. 2017) and SegNet (Badrinarayanan, Kendall, and Cipolla 2017) are evaluated on EndoVis 2017. The results of other methods are all from EndoVis challenge 2017 (Allan et al. 2019).

Results show that our method achieves 64.10% mean IOU, outperforming existing methods by a large margin. The second-ranking method is TernausNet which achieves 54.20% mean IOU. Compared with TernausNet, our network has increased by 9.90% on mean IOU, which is a significant margin. Besides, the segmentation results of each video sequence are shown in Table 4. Our method achieves the best results in 7 video sequences. On the other three

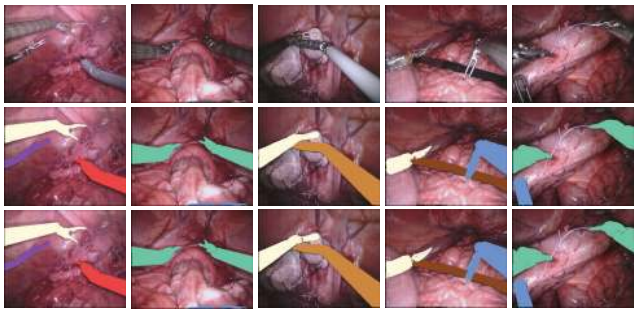

Figure 7: Visualization for segmentation results of PAANet on EndoVis 2017. From top to bottom: image, prediction and ground truth.

video sequences, the performance of the proposed network is also at the forefront. These results show that our network achieves state-of-the-art performance on this dataset.

To give a more intuitive display, the segmentation results of PAANet are visualized in Figure 7. Despite the dramatic changes in the scale and shape of surgical instruments, the proposed network can still segment them correctly. Predictions are the same as the ground truth, proving the excellent performance of the proposed network.

Conclusion

In this paper, we propose the Pyramid Attention Aggregation Network to address specular reflection and scale variation issues, which includes two critical modules: the Double Attention Module and the Pyramid Upsampling Module. Ablation experiments show that the Double Attention Module can improve segmentation accuracy by modeling semantic dependencies. Experiments also prove that the Pyramid Upsampling Module can significantly improve segmentation accuracy with very few parameters. Besides, the proposed network achieves state-of-the-art performance on Cata7 and EndoVis 2017.

Acknowledgments

This research is supported by the National Key Research and Development Program of China (Grant 2017YFB1302704), the National Natural Science Foundation of China (Grants 61533016, U1713220), the Beijing Science and Technology Plan(Grant Z191100002019013) and 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.; and Azizian, M. 2019. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*.

Attia, M.; Hossny, M.; Nahavandi, S.; and 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)*, 3373–3378. IEEE.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39(12):2481–2495.

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

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

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4):834–848.

Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; and Feng, J. 2018b. Aˆ2-nets: Double attention networks. In *Advances in Neural Information Processing Systems 31*. 352–361.

Fu, J.; Liu, J.; Tian, H.; Fang, Z.; and Lu, H. 2018. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*.

García-Peraza-Herrera, L. C.; Li, W.; Gruijthuijsen, C.; Devreker, A.; Attilakos, G.; Deprest, J.; Vander Poorten, E.; Stoyanov, D.; Vercauteren, T.; and Ourselin, S. 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In *International Workshop on Computer-Assisted and Robotic Endoscopy*, 84–95. Springer.

García-Peraza-Herrera, L. C.; Li, W.; Fidon, L.; Gruijthuijsen, C.; Devreker, A.; Attilakos, G.; Deprest, J.; Vander Poorten, E.; Stoyanov, D.; Vercauteren, T.; et al. 2017. Toolnet: Holistically-nested real-time segmentation of robotic surgical tools. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5717–5722. IEEE.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

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

Jin, Y.; Cheng, K.; Dou, Q.; and 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.; and Khan, A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 440–448. Cham: Springer International Publishing.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and

Zhang, B.-T. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

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

Laina, I.; Rieke, N.; Rupprecht, C.; Vizcaíno, J. P.; Eslami, A.; Tombari, F.; and Navab, N. 2017. Concurrent segmentation and localization for tracking of surgical instruments. In *International conference on medical image computing and computer-assisted intervention*, 664–672. Springer.

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

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

Qin, F.; Li, Y.; Su, Y.; Xu, D.; and 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)*, 9821–9827.

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

Sarikaya, D.; Corso, J. J.; and 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.

Shvets, A. A.; Rakhlin, A.; Kalinin, A. A.; and Iglovikov, V. I. 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 624–628.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Yang, M.; Yu, K.; Zhang, C.; Li, Z.; and Yang, K. 2018. Denseaspp for semantic segmentation in street scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3684–3692.

Yu, Z.; Yu, J.; Fan, J.; and 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)*, 1821–1830.

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

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230–6239.