Space Squeeze Reasoning and Low-Rank Bilinear Feature Fusion for Surgical Image Segmentation

Zhen-Liang Ni, Gui-Bin Bian, *Member, IEEE*, Zhen Li, Xiao-Hu Zhou, *Member, IEEE*, Rui-Qi Li, and Zeng-Guang Hou, *Fellow, IEEE*

***Abstract*—Surgical image segmentation is critical for surgical robot control and computer-assisted surgery. In the surgical scene, the local features of objects are highly similar, and the illumination interference is strong, which makes surgical image segmentation challenging. To address the above issues, a bilinear squeeze reasoning network is proposed for surgical image segmentation. In it, the space squeeze reasoning module is proposed, which adopts height pooling and width pooling to squeeze global contexts in the vertical and horizontal directions, respectively. The similarity between each horizontal position and each vertical position is calculated to encode long-range semantic dependencies and establish the affinity matrix. The feature maps are also squeezed from both the vertical and horizontal directions to model channel relations. Guided by channel relations, the affinity matrix is expanded to the same size as the input features. It captures long-range semantic dependencies from different directions, helping address the local similarity issue. Besides, a low-rank bilinear fusion module is proposed to enhance the model's ability to recognize similar features. This module is based on the low-rank bilinear model to capture the inter-layer feature relations. It integrates the location details from low-level features and semantic information from high-level features. Various semantics can be represented more accurately, which effectively improves feature representation. The proposed network achieves state-of-the-art performance on cataract image segmentation dataset CataSeg and robotic image segmentation dataset EndoVis 2018.**

***Index Terms*—Bilinear feature fusion, space squeeze reasoning, surgical image segmentation.**

I. Introduction

SURGICAL image segmentation is a critical technology for computer-assisted surgery and surgical robot navigation [1]–[5]. Its goal is to segment objects such as biological tissues and surgical instruments in the surgical scene and assign a category label to each pixel. The segmentation results can be utilized in clinical work, such as diseased tissue localization, surgical instrument tracking, and visual enhancement [1]. Besides, this technology automates many postoperative tasks, including surgical report generation, objective assessment of skills, surgical video retrieval, and so on [1]. These applications can reduce the workload of doctors and improve the success rate of surgery, which is beneficial for clinical work [6].

Surgical image segmentation is a challenging task, which faces many difficulties. The most important issue is the high similarity of local features. Different biological tissues often have similar visual features such as color and texture in local areas. As shown in Fig. 1(a), the iris and pupil in the cataract scene are all brown and the biological tissues in the endoscopic scene are all red. Besides, the shapes of various surgical instruments are very similar in local regions. As illustrated in Fig. 1(a), the cataract surgery instruments are all slender in local regions and their colors are very similar. Thus, it is difficult to distinguish them based on local features. Furthermore, there are some illumination interferences in the surgical scene. Illumination interference changes the color and texture of the object. As shown in Fig. 1(b), specular reflections cause objects to become bright white and shadows cause objects to become black. These interferences also cause the similarity of local features. In recent years, some work on surgical image segmentation has been proposed. TDSNet [7] adopted a task decomposition strategy to balance the deviation

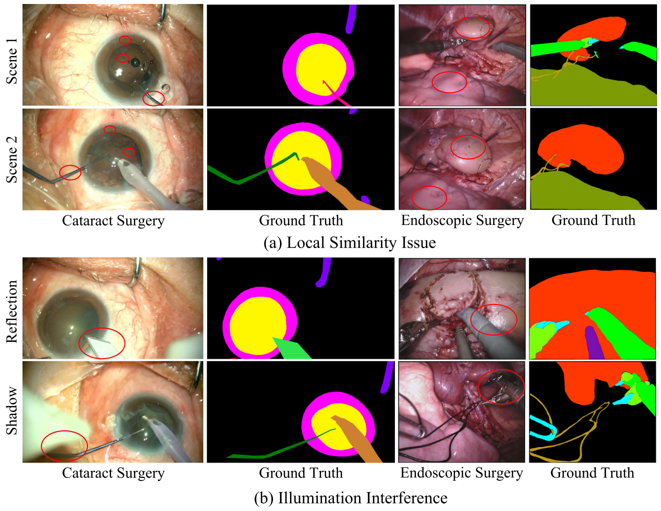
Manuscript received September 15, 2021; revised December 3, 2021 and January 25, 2022; accepted February 19, 2022. Date of publication February 28, 2022; date of current version July 4, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 62027813, U20A20196, and U1713220, in part by the National Key Research and Development Program of China under Grant 2020YFF01014800ZL, in part by the CAS Interdisciplinary Innovation Team under Grant JCTD-2019-07, in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant 2018165, and in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB32040000. *(Corresponding authors: Gui-Bin Bian; Zeng-Guang Hou.)*

Zhen-Liang Ni, Gui-Bin Bian, and Rui-Qi Li are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: nizhenliang2017@ia.ac.cn; guibin.bian@ia.ac.cn; liruiqi2016@ia.ac.cn).

Zhen Li and Xiao-Hu Zhou are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhen.li@ia.ac.cn; xiaohu.zhou@ia.ac.cn).

Zeng-Guang Hou is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, CAS Center for Excellence in Brain Science and Technology, Chinese Academy of Sciences, Beijing 100190, China, and with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS-MUST Joint Laboratory of Intelligence Science and Technology, Institute of Systems Engineering, Macau University of Science and Technology, Macau 999078, China (e-mail: zengguang.hou@ia.ac.cn).

Digital Object Identifier 10.1109/JBHI.2022.3154925



(a) Local Similarity Issue

(b) Illumination Interference

Fig. 1. Some difficult samples in the surgical scene. The surgical scene mainly includes biological tissues and surgical instruments. (a) Local similarity Issue: Some biological tissues have similar visual features in local regions, which are marked by red circles. (b) Illumination Interference: specular reflection makes objects tend to be white and shadows make objects tend to be black, causing the visual features of different objects to become similar.

between the pixel-wise semantic segmentation and the instance class prediction tasks. Spatio-Temporal Multi-Task Learning (ST-MTL) [8] is an end-to-end trainable model for real-time surgical instrument segmentation and task-oriented saliency detection, which consists of a shared encoder and spatio-temporal decoders. A general embeddable approach [9] introduced the multi-angle feature aggregation (MAFA) method to adapt to instrument orientation variation. DeepLabV3+ [10] is applied to surgical image segmentation in the 2018 EndoVis Robotic Scene Segmentation Challenge and takes the first place [2]. BARNet [11] designed an adaptive receptive field module to cope with the scale variation of surgical instruments. However, most methods only mainly focus on surgical instrument segmentation and ignore the biological tissues. The segmentation of biological tissues is also crucial for computer-assisted surgery. Besides, these methods improve segmentation accuracy in different ways but do not focus on the local similarity issue, limiting their performance.

To address the local similarity issue, we consider two aspects. First, long-range semantic dependencies can make the network learn the overall shape features to distinguish local similar features. The semantic dependencies can also be used to infer the features in the illumination interference area from the neighboring pixels. Second, the fusion of high-level features and low-level features can obtain more comprehensive information and enhance the distinction between similar features.

The space squeeze reasoning module is proposed to model long-range semantic dependencies for distinguishing local-similar features. This module adopts height pooling and width pooling to squeeze global contexts in the vertical and horizontal directions, respectively. Then, the similarity between each horizontal position and each vertical position is calculated to encode long-range semantic dependencies and establish an affinity matrix. Besides, the feature maps are squeezed from both the vertical and horizontal directions to model the semantic relations between channels. Guided by channel relations, the affinity matrix is adaptively distributed to the original feature space. In this way, the space squeeze reasoning module can capture long-range semantic dependencies from different directions, addressing the local similarity issue.

The low-rank bilinear fusion module is proposed to fuse high-level features and low-level features effectively. It is based on low-rank bilinear pooling to capture relations of different level features. Specifically, the input two-level features are embedded in the new feature spaces. Hadamard product is applied for cross-layer feature interaction, which helps to enhance the distinction between different semantic features. The merged features integrate the location details in the low-level feature maps and the semantic information in the high-level feature maps, which can more accurately represent various semantics. Furthermore, we apply channel attention to further improve the feature representation. In this way, the low-rank bilinear fusion module can enhance the ability to recognize similar features. Based on the above analysis, a bilinear squeeze reasoning network (SRBNet) including space squeeze reasoning and low-rank bilinear fusion is proposed for surgical image segmentation. The contributions of this work can be summarized as follows:

- The space squeeze reasoning module is proposed to squeeze feature maps from different directions for encoding multi-directional long-range semantic dependencies, helping to address local similarity issues.
- The low-rank bilinear fusion module is proposed to fuse low-level and high-level features. It integrates features of different levels to enhance the distinction between similar features.
- The proposed network achieves state-of-the-art performance on CataSeg and gets a new record on EndoVis 2018.

II. Related Work

A. Surgical Image Segmentation

Recently, most of the work related to surgical image segmentation focuses on the segmentation of surgical instruments. Some methods improved segmentation accuracy by capturing shape priors of instruments. For example, ToolNet-C combined with the kinematic pose information to get the accurate silhouette mask [16]. MF-TAPNet [17] adopted optical flow as prior to provide a reliable indication of the instrument location and shape for accurate segmentation. These methods require additional information to assist the segmentation, which is not conducive to the deployment of the model.

Other methods introduced various modules to improve feature representations. For instance, RAUNet [18] designed an attention module to fuse multi-level feature maps and emphasize the target region. A hybrid CNN-RNN method [19] introduced the Recurrent Neural Network to capture global contexts and expand the receptive field. Spatio-Temporal Multi-Task Learning (ST-MTL) [8] is an end-to-end trainable model for real-time surgical instrument segmentation and task-oriented saliency detection, which consists of a shared encoder and spatio-temporal decoders. Auxiliary Supervised Deep Adversarial Learning (ASDAL) [20] is proposed to regularize the segmentation model. In this work, auxiliary supervision helps the model to learn low-resolution features, and adversarial learning improves the segmentation prediction by learning higher-order structural information. The Multi-Angle Feature Aggregation (MAFA) method [9] is proposed, which leverages active image rotation to gain richer visual cues and make the prediction more robust to instrument orientation changes. However, these methods cannot identify and segment biological tissues, limiting their application scenarios. The proposed method takes into account both surgical instruments and biological tissues, thus it has a wider range of applications. TDSNet [7] adopted a task decomposition strategy to balance the deviation between the pixel-wise semantic segmentation and the instance class prediction tasks. But, it does not consider the similarity of local features in surgical scenes, which limits their performance.

Besides, the 2020 CATARACTS Semantic Segmentation Challenge [1] and the 2018 EndoVis Robotic Scene Segmentation Challenge [2] are held to make researchers pay attention to surgical image segmentation. They also make the challenge dataset public, which is helpful for other researchers to develop deep learning methods in this direction.

B. Attention Model in Semantic Segmentation

A series of attention modules are applied in semantic segmentation [12], [21]–[23]. Squeeze and excitation block [12] applied global average pooling to capture global contexts and model semantic dependencies between channels. Coordinate attention [13] introduced coordinate information to model semantic dependencies. However, the squeeze and excitation block directly squeezes the feature map to the vector representation, losing the spatial information. Coordinate attention only captures the spatial coordinate information without paying attention to the relationship between channels, which limits its performance. The proposed SSRM not only captures the spatial semantic dependence relationship but also models the relationship between channels. Since it captures multiple forms of information, SSRM is more conducive to improving feature representation. Non-local network [14] captured long-range semantic relationships, making feature representation global. Dual attention network [15] captured global contexts and semantic dependencies between channels to improve performance. However, the non-local block [14] and dual attention network [15] calculate the similarity between all pixels and thus they need high computational cost. The proposed SSRM uses width and height pooling to pool the feature map, which reduces the computational cost.

C. Bilinear Model

Bilinear models are widely used in computer vision tasks. The bilinear CNN [24] was proposed to model local pairwise feature relationships in a translation-invariant manner for fine-grained visual recognition. A2Net [21] adopted bilinear pooling to model semantic dependencies and generate attention features. Bilinear models can capture second-order statistics and improve feature representation. However, bilinear models often require high computational costs. To reduce computational costs, low-rank bilinear pooling [25] factorized the bilinear model by using the Hadamard product. Based on low-rank bilinear pooling, hierarchical bilinear pooling [26] was proposed to fuse multiple cross-layer bilinear features and capture the inter-layer feature relations. However, low-rank bilinear models have rarely been applied to semantic segmentation. The semantic segmentation task also needs to capture fine-grained features for the correct recognition of object categories. Thus, we applied the low-rank bilinear model in SRBNet to learn fine-grained features and improve segmentation accuracy.

III. Methodology

A. Overview

Local features of various biological tissues or surgical instruments are often similar, making the segmentation challenging. Long-range semantic dependencies are essential to segment similar objects, which can help the network learn overall shape features to distinguish them. Furthermore, specular reflections and shadows will change the color and texture of objects, reducing the contrast between objects. It is difficult for the network to recognize objects based on these visual features. To address these issues, we propose the space squeeze reasoning module (SSRM) to capture long-range semantic dependencies and adaptively distribute global features based on channel relationships. The low-rank bilinear fusion module (LBFM) is proposed to learn fine-grained features and enhance the ability to recognize visual features.

The architecture of the proposed SRBNet is shown in Fig. 2. It adopts encoder-decoder architecture. The dilated ResNet [27] is used as the backbone. Dilated convolution can expand the receptive field and preserve location details. SSRM is used to model global contexts in the output feature of the encoder. Three LBFMs are designed to fuse low-level feature maps and high-level feature maps. To reduce the computational costs, the final output is upsampled four times to obtain a high-resolution mask.

B. Space Squeeze Reasoning Module

The long-range semantic dependencies are essential to learning the overall shape of objects, contributing to addressing the local feature similarity issue. The spatial squeeze reasoning module models long-range semantic dependencies from different directions, which not only captures the location information but also the semantic relationship between channels.

1) Space Squeezing and Reasoning: SSRM first squeezes the input from the vertical and horizontal directions to obtain a pair of direction-aware features. First, 1×1 convolution with nonlinear activation and space pooling are applied for feature embedding. Height pooling and width pooling are adopted to aggregate location information in the vertical and horizontal directions, respectively. Specifically, the height pooling squeezes feature maps along the vertical direction and generates $y^h \in \mathbb{R}^{C\times W}$.

$$y^h(k,j) = g_h(x) = \frac{1}{H}\sum_{i=1}^{H} x(k,i,j) \tag{1}$$

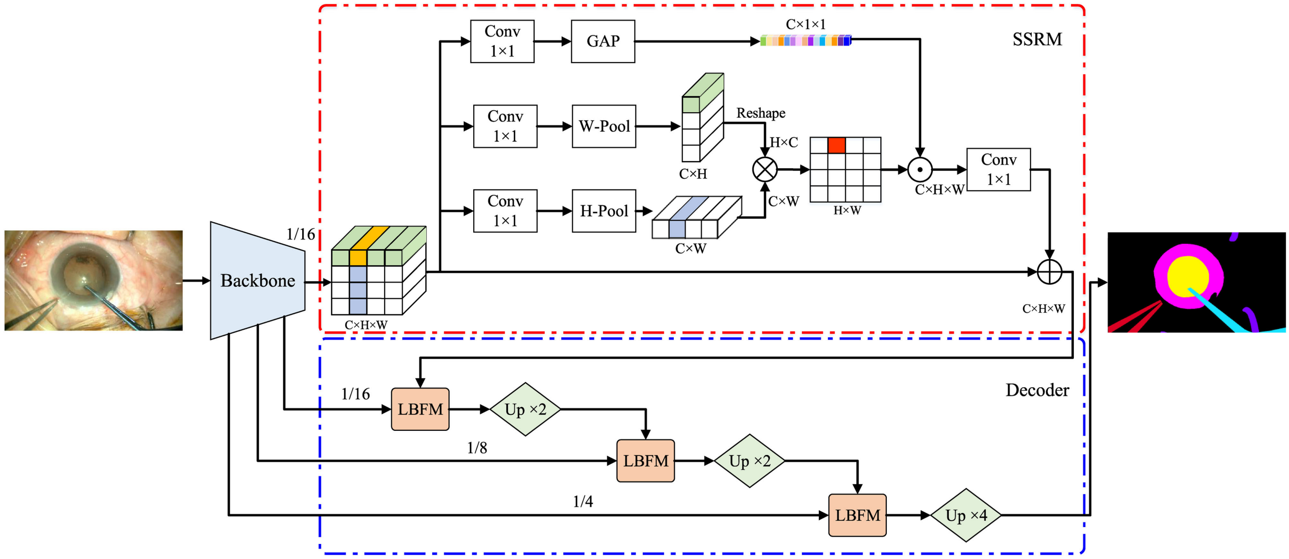


Fig. 2. The architecture of the SRBNet. The space squeeze reasoning module (SSRM) is designed to capture long-range semantic dependencies. The low-rank fusion module (LBFM) is proposed to fuse different-level features and learn fine-grained features. Dilated ResNet is used as the backbone. ⊗ represents matrix multiplication. ⊙ denotes broadcast Hadamard product. ⊕ refers to element-wise addition.

The width pooling squeezes feature maps along horizontal direction and generates $y^w \in \mathbb{R}^{C\times H}$.

$$y^w(k,i) = g_w(x) = \frac{1}{W}\sum_{j=1}^{W} x(k,i,j) \tag{2}$$

where $x \in \mathbb{R}^{C\times H\times W}$ is the input feature map. k, i, j are the indexes of the channel, height and width of the input, respectively. $k \in \{1,2,...,C\}, i \in \{1,2,...,H\}, j \in \{1,2,...,W\}$.

Then, matrix multiplication is performed on the paired direction-aware features to get an affinity matrix. The affinity matrix encodes long-range semantic dependencies by calculating the similarity between each horizontal position and each vertical position. The information flow of squeeze reasoning operations is shown in Fig. 2. The red pixel is calculated from all the elements in the row and column where the orange box is located. Therefore, each pixel of the affinity matrix can capture location information on corresponding rows as well as columns and integrates global contexts along the channel direction.

$$\overline{y} = g_w(W_\theta x) \times g_h(W_\phi x) \tag{3}$$

where $\overline{y} \in \mathbb{R}^{H\times W}$ is the affinity matrix. g_w represents width pooling and g_h represents height pooling. W_θ and W_ϕ refer to 1×1 convolution with ReLU.

In summary, the squeeze inference operation can not only capture the long-range dependencies in space but also retain the original location information, which is helpful for the positioning of the target object. Due to the height and width pooling, the computational costs of matrix multiplication are very low.

2) Feature Embedding: The affinity matrix encoding long-range semantic dependencies need to be embedded in the original feature space. The global average pooling is adopted to distribute the affinity matrix to the original feature space, which is illustrated in (4). It squeezes feature maps along both the vertical and horizontal directions into a channel relation vector, which can capture global contexts in each element. This vector is fed into convolution and nonlinear activation encodes the semantic relationships along channels.

$$y^c(k) = g_c(x) = \frac{1}{W\times H}\sum_{i=1}^{H}\sum_{j=1}^{W} x(i,j,k) \tag{4}$$

where $y^c \in \mathbb{R}^{C\times 1\times 1}$ is the attention vector.

The broadcast Hadamard product is applied to weight the affinity matrix with the channel relation vector obtained by global average pooling. The affinity matrix is extended to the same dimensions as the channel relation vector by the broadcast Hadamard product. Moreover, the long-range semantic dependencies are embedded in each channel. The expanded features undergo 1×1 convolution and nonlinear activation to further refine the features. Moreover, residual connections are added to calibrate semantic features, helping to improve feature representation.

$$y = \sigma\left(f\left(\overline{y} \odot y^c\right)\right) + x \tag{5}$$

where $y \in \mathbb{R}^{C\times H\times W}$ represents the output feature map. ⊙ denotes broadcast Hadamard product. f refers to 1×1 convolution. σ represents the ReLU. This feature embedding operation squeezes the global features of each channel to model the semantic relationships between channels, improving the feature representation. The time complexity of SSRM and the non-local block are compared. The proposed SSRM adopts width pooling and height pooling to squeeze feature maps, which can significantly reduce the time complexity. The time complexity of matrix multiplication in SSRM is only $\mathcal{O}(CHW)$, which is significantly lower than $\mathcal{O}(CH^2W^2)$ in the standard non-local block. Besides, in the feature embedding stage, the time complexity of the broadcast Hadamard product in SSRM is also $\mathcal{O}(CHW)$. And, the non-local block still uses matrix multiplication with the time complexity of $\mathcal{O}(CH^2W^2)$. Thus, the time complexity of the proposed SSRM is significantly lower than the non-local block.

C. Low-Rank Bilinear Fusion Module

The fusion of high-level features and low-level features can obtain more comprehensive information and enhance the distinction between similar features. Thus, the low-rank bilinear fusion module is proposed to fuse high-level features and low-level features effectively. LBFM fuses the low-level feature map from the encoder and the high-level feature map from the decoder. It is based on the low-rank bilinear model to capture the inter-layer feature relations, which is an important technique for fine-grained recognition [26]. Low-rank bilinear model using Hadamard product to factorize the bilinear model [25], which reduces computational costs and retains the ability to extract fine-grained features. The low-rank bilinear is shown in (6).

$$z_{ij} = U^T x_{ij} \circ V^T y_{ij} \tag{6}$$

where $x \in \mathbb{R}^{C \times H \times W}$ represents high-level input feature maps and $y \in \mathbb{R}^{C \times H \times W}$ represents low-level feature maps. $x_{ij}, y_{ij} \in \mathbb{R}^{C \times 1}$, where $i = 1, 2, ..., H$ and $j = 1, 2, ..., W$. $U \in \mathbb{R}^{C \times C}$ and $V \in \mathbb{R}^{C \times C}$ denote projection matrixes. $z_{ij} \in \mathbb{R}^{C \times 1}$ is the output feature. $\circ$ refers to Hadamard product. In LBFM, the projection matrix is replaced by convolution with kernel size 1×1.

Specifically, convolutions are first applied to embed two feature layers into the new feature space. Then, the Hadamard product is applied to calculate the relationship of features between layers. Compared with addition, the Hadamard product can enhance the distinction between semantic various features and effectively improve the feature representation. Since the low-level feature map contains rich location details and the high-level feature map contains rich semantic features, the fused features can represent various semantics more accurately. In this way, the capacity of the model to segment similar objects can be improved.

Besides, we exploit channel attention to further improve the feature representation [12]. Specifically, global average pooling is used to compress feature maps, obtaining a weight vector. And two layers of convolution are applied to model channel relationships [12]. Then, the weight vector is weighted to the fused feature. Finally, the high-level feature and the fused feature are added to preserve the semantic information as much as possible, which also helps to reduce the loss of semantic information in the propagation process.

$$\overline{z} = \phi(z) + x \tag{7}$$

where $\overline{z} \in \mathbb{R}^{C \times H \times W}$ represents the final outputs of low-rank bilinear fusion module. ϕ refers to the channel attention.

IV. Experiments

A. Dataset

1) CataSeg: CataSeg records the scene of cataract surgery. It contains 2236 images from 7 cataract surgery videos. The resolution of each image is 1920×1080. There are 13 types of objects in the scenes, including 11 surgical instrument classes and 2 biological tissue. The dataset is split into a training set and a test set. 1416 images from videos 2, 4, 5, and 7 are used for training. 820 images from videos 1, 3, and 6 are used for testing.

2) EndoVis 2018: EndoVis 2018 is from the MICCAI Endovis Robotic Scene Segmentation Challenge 2018 [2], which is based on endoscopic surgery. This dataset is acquired from the da Vinci Xi robot. The training set contains 15 video sequences with a total of 2235 images. Each sequence contains 149 images with a resolution of 1280×1024. The test set contains 1000 images. Similar frames were manually removed. There are 11 types of objects in this dataset.

B. Implementation Details

All experiments are implemented based on PyTorch, using GPU Tesla V100. The dilated ResNet [27] is used as the backbone. The dilation rate is set to 2 in the last stage of the backbone. Adam with default parameters of PyTorch is adopted as the optimizer and the batch size is 8. To prevent over-fitting, the learning rate is dynamically adjusted during training. Specifically, for every 30 iterations, the learning rate is multiplied by 0.8. The initial learning rates of CataSeg dataset is 4×10^{-5} and that of the EndoVis 2018 dataset is 3.5×10^{-5}. All images in EndoVis 2018 are resized to 640×512. The size of images in CataSeg is resized to 640×384. The focal loss [28] is adopted to train the proposed method, dealing with the class imbalance issue, which is described in (8). The models used for comparison are trained with the same settings except that the initial learning rate is different.

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t), \gamma \geq 0 \tag{8}$$

where γ is used to reduce the weight of easy samples, making the model more focused on hard samples during training. The γ is set to 4 in all experiments.

To objectively evaluate the proposed method, the Dice coefficient and Intersection-over-Union (IoU) are selected as the evaluation metric. They are used to evaluate the similarity between ground truth and predictions.

$$Dice = \frac{2\,|G \cap P|}{|G| + |P|}, IoU = \frac{|G \cap P|}{|G \cup P|} \tag{9}$$

where G refers to the ground truth and P refers to the prediction result.

The mean values of IoU and Dice are represented as mIoU and mDice, respectively.

$$mDice = \frac{1}{n}\sum_{i}^{n} d_i, mIoU = \frac{1}{n}\sum_{i}^{n} u_i \tag{10}$$

where d_i and u_i represent the Dice and IoU of the i-th category, respectively. n is the total number of categories.

C. CataSeg

1) Ablation Study for Space Squeeze Reasoning Module: The space squeeze reasoning module(SSRM) is designed to capture long-range semantic dependencies, helping to address the local similarity issue. To verify its performance, some experiments are performed, which are shown in Table II.

TABLE I
COMPARISON OF THE ADVANTAGES AND DISADVANTAGES OF DIFFERENT ATTENTION METHODS

Method	Advantage	Disadvantage
Squeeze and excitation block [12]	Capture channel relation	No spatial information
Coordinate attention [13]	Capture spatial coordinate information	No channel relation
Non-local network [14]	Capture spatial global contexts	No channel relation High computational complexity
Dual attention network [15]	Capture spatial relation and channel relation	High computational complexity
SSRM(Ours)	Capture spatial relation and channel relation Low computational complexity	

TABLE II
QUANTITATIVE PERFORMANCE ANALYSIS OF SPATIAL SQUEEZE REASONING MODULE

Method	Blocks	mDice(%)	mIoU(%)
Baseline	None	85.48	77.53
Baseline	Non-Local [14]	87.60	79.71
Baseline	CA [13]	87.55	80.16
Baseline	SSRM	89.50	82.42

'CA' represents coordinate attention.

TABLE III
ABLATION EXPERIMENTS FOR THE LOW-RANK BILINEAR FUSION MODULE

Methhod	SSRM	LBFM	mDice(%)	mIoU(%)
Baseline1	×	×	84.76	76.32
Baseline1	×	3	85.48	77.53
Baseline1	1	×	87.42	79.47
Baseline1	1	1	87.73	80.35
Baseline1	1	2	89.42	81.81
Baseline1	1	3	89.50	82.42

× means that the module is not used.

TABLE IV
PERFORMANCE COMPARISON OF VARIOUS METHODS ON CATASEG. FPS IS CALCULATED ON TITAN X WITH A BATCH SIZE OF 1

Method	Backbone	mDice(%)	mIoU(%)	FPS
LinkNet [29]	ResNet50	80.92	71.87	19.43
RAUNet [18]	ResNet50	80.21	70.71	26.31
RefineNet [30]	ResNet50	81.13	72.44	16.56
PAN [23]	ResNet50	85.16	76.55	26.30
BARNet [11]	ResNet50	86.28	78.31	19.01
PSPNet [31]	ResNet50	87.16	79.86	15.91
DeepLabV3+ [10]	ResNet50	86.87	79.38	25.69
SRBNet(Ours)	ResNet50	89.50	82.42	26.04

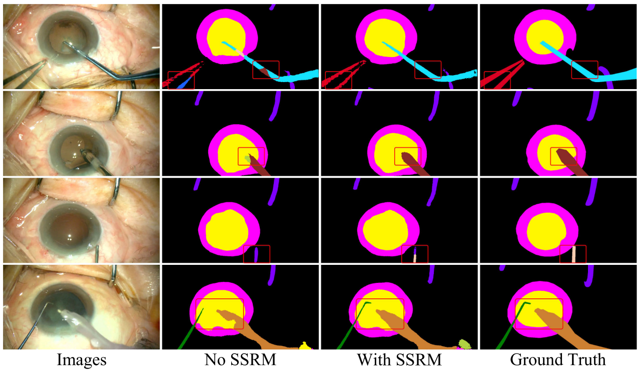


Fig. 3. Qualitatively analysis for the performance of the space squeeze reasoning module on CataSeg. Various objects are represented by different colors. By applying SSRM, the recognition errors are significantly reduced.

SRBNet without SSRM is used as the baseline network in Table II, which achieves 85.48% mean Dice and 77.53% mean IoU. Compared with the baseline network, the network adding SSRM brings 4.02% mean Dice and 4.89% mean IoU growth due to capturing the long-range semantic dependencies. To further prove its excellent performance, non-local and coordinate attention instead of SSRM are tested for comparison. It can be observed that the baseline network with non-local block achieves 87.60% mean Dice and 79.71% mean IoU. Compared with it, the mean Dice and mean IoU of the baseline network with SSRM increase by 1.90% and 2.71%, respectively. The baseline network with coordinate attention achieves 87.55% mean Dice and 80.16% mean IoU, which reduces mean Dice by 1.95% and mean IoU by 2.26% compared to SSRM. The above results prove that SSRM can significantly improve segmentation performance.

Besides, segmentation results of baseline and baseline with SSRM are visualized to qualitatively analyze the effect of SSRM. As shown in Fig. 3, the segmentation results with SSRM have fewer recognition errors compared to predictions without SSRM. The visual results of applying SSRM are also more complete and closer to the ground truth. These visual results show that SSRM can effectively distinguish similar features.

2) Ablation Study for Low-Rank Bilinear Fusion Module: Low-rank bilinear fusion module (LBFM) is proposed to learn fine-grained features and enhance the discrimination for similar features. To evaluate its performance, the ablation experiment is performed. The results are illustrated in Table III.

The baseline1 refers to the SRBNet without SSRM and LBFM, which achieves 84.76% mean Dice and 76.32% mean IoU. Compared with the baseline1, employing LBFM brings 0.72% mean Dice and 1.21% mean IoU. Besides, LFBM can further increase the mean Dice by 2.08% and mean IoU by 2.95% on the basis of the baseline with SSRM. Furthermore, we test the impact of the number of LFBM on network performance. It can be observed that the performance is best when three LFBM are used. These results prove the effectiveness of LBFM.

3) Comparison With State-of-the-Arts: To evaluate the performance of SRBNet, a series of state-of-the-art methods are tested on the CataSeg. As shown in Table IV, the SRBNet achieves state-of-the-art performance 89.50% mean Dice and 82.42% mean IoU, which outperforms the second method PSPNet [31] by 2.34% on mean Dice and 2.56% on mean IoU.

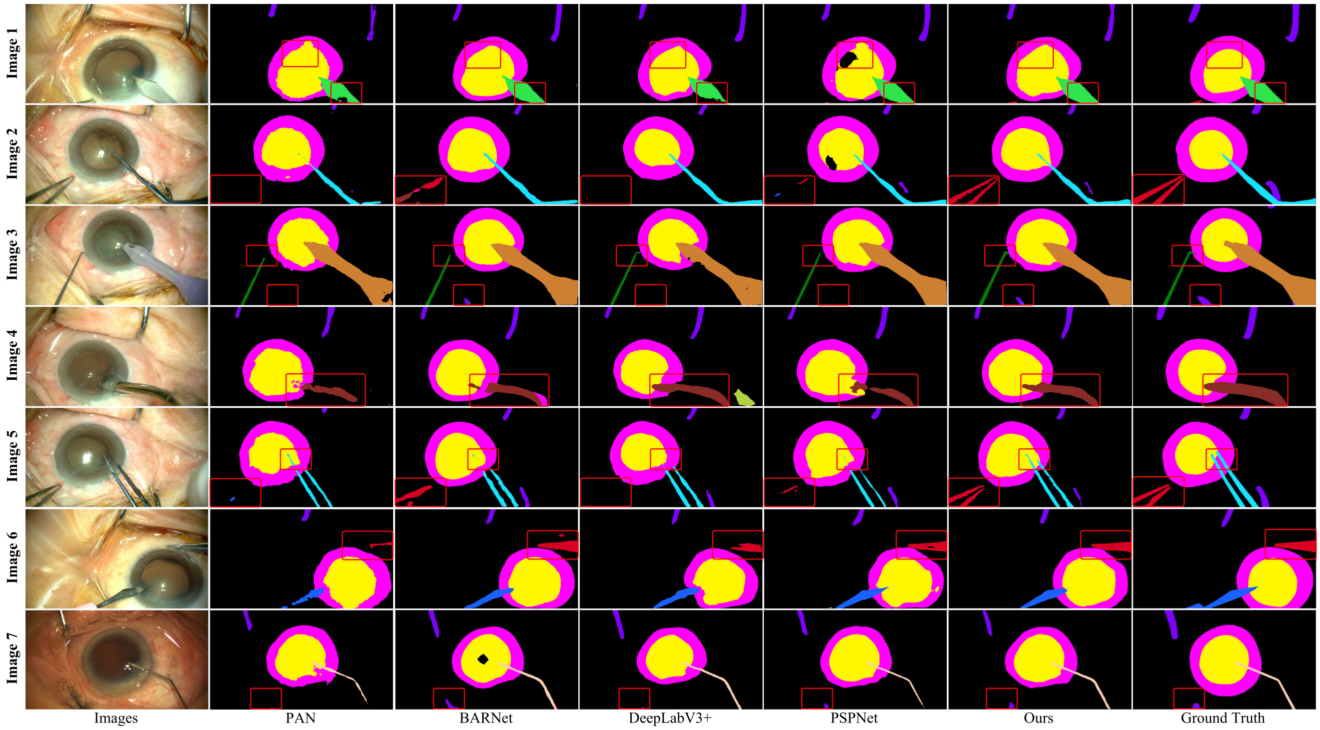


Fig. 4. Visualization of segmentation results by different methods. The rectangles mark the main contrast area. Various objects are represented by different colors. Even if the colors of different biological tissues are very similar, the proposed SRBNet can still segment them well and the segmentation results are similar to the ground truth.

DeepLabV3+ [10] have got the best result in the MICCAI Endovis Robotic Scene Segmentation Challenge 2018 and it is also widely used in surgical image segmentation tasks [32]. SRBNet exceeds DeepLabV3+ [10] by 2.63% mean Dice and 3.04% mean IoU. Besides, SRBNet also exceeds BARNet [11] by 3.22% mean Dice and 4.11% mean IoU. The other methods are much poorer than the proposed method. Further, the inference speed of the model is calculated and shown in Table IV. FPS is used as the evaluation metric. FPS is calculated on Titan X with a batch size of 1. The inference speed of SRBNet is 26 fps, which is faster than most comparable models. The inference speed of SRBNet is slightly slower than RAUNet and PAN, but the accuracy far exceeds them.

4) Mask Visualization: To give more intuitive results, the segmentation results of various methods and the proposed SRBNet are visualized in Fig. 4.

In images 1, 3, and 6, the visual features of the iris and pupil are very similar, making it difficult to distinguish their boundaries. It can be found that SRBNet can segment the iris and pupil more accurately than other methods. This is because SSRM captures long-range semantic dependence to distinguish local similar features. In image 4, the segmentation results of PSPNet, DeepLabV3+, and BARNet all have obvious recognition errors. Compared with them, the proposed SRBNet has fewer recognition errors. In images 2, 5, and 7, other methods can not completely segment some surgical instruments. Our segmentation results are more complete and closer to the ground truth. The above visualization results prove the excellent performance of SRBNet qualitatively.

TABLE V
PERFORMANCE COMPARISON OF VARIOUS METHODS ON ENDOVIS 2018

Method	Backbone	mDice(%)	mIoU(%)
Unet [33]	ResNet50	43.95	34.83
RAUNet [18]	ResNet50	68.89	58.78
RefineNet [30]	ResNet50	69.05	58.90
PAN [23]	ResNet50	68.91	59.32
BARNet [11]	ResNet50	70.10	59.92
DeepLabV3+ [10]	ResNet50	70.69	60.94
SRBNet(Ours)	ResNet50	71.90	62.19

The proposed SRBNet achieves the best performance, exceeding the second-place method by 1.21% on mean dice and 1.25% on mean IoU.

D. EndoVis 2018

1) Comparison With State-of-the-Arts: The proposed method is also evaluated on EndoVis 2018 to further verify its performance. A series of excellent methods are regarded as comparison methods. All results are shown in Table V. The SRBNet achieves the best performance on EndoVis 2018, which achieves 71.90% mean Dice and 62.19% mean IoU. The second-ranking method, DeepLabV3+, achieves 70.69% mean Dice and 60.94% mean IoU. Its mean Dice and mean IoU are 1.21% and 1.25% lower than that of SRBNet, respectively. Besides, the BARNet gets 70.10% mean Dice and 59.92% mean IoU, whose mean Dice and mean IoU are 1.80% and 2.27% lower than that of SRBNet. The performance of other methods is much poorer than SRBNet. The above results prove that the proposed method achieves state-of-the-art performance.

TABLE VI
MIOU AND MDICE OF EACH CATEGORY ON ENDOVIS 2018

Method	Metric	Shaft	Clasper	Wrist	Parenchyma	Kidney	Thread	Clamps	Needle	Intestine	US_probe
Unet	mIoU	88.35	40.79	43.78	62.54	16.60	0.09	0.14	0.00	74.22	21.76
	mDice	93.81	57.94	60.90	76.95	28.47	0.18	0.28	0.00	85.20	35.74
RAUNet	mIoU	93.19	56.35	63.38	**92.58**	**66.88**	29.93	52.17	0.00	96.98	**36.38**
	mDice	96.47	72.08	77.58	**96.15**	**80.16**	46.07	68.57	0.00	98.46	**53.35**
RefineNet	mIoU	93.21	59.47	65.39	90.76	56.51	36.80	66.97	0.00	92.65	27.22
	mDice	96.48	74.58	79.07	95.16	72.21	53.80	80.22	0.00	96.18	42.80
PAN	mIoU	92.75	59.69	66.54	92.25	60.04	34.59	**72.23**	0.00	93.76	21.32
	mDice	96.24	74.76	79.91	95.97	75.03	51.40	**83.88**	0.00	96.78	35.15
DeepLabV3+	mIoU	93.56	58.20	68.68	90.60	61.87	35.80	70.27	0.00	97.31	33.14
	mDice	96.67	73.58	81.43	95.07	76.44	52.73	82.54	0.00	98.64	49.79
BARNet	mIoU	92.97	60.72	66.90	91.05	57.65	39.81	59.10	**0.16**	97.21	33.66
	mDice	96.36	75.56	80.17	95.31	73.13	56.95	74.29	**0.32**	98.59	50.37
NLBNet(Ours)	mIoU	**93.91**	**62.12**	**70.02**	90.88	60.19	**48.24**	65.06	0.00	**98.51**	32.99
	mDice	**96.86**	**76.64**	**82.37**	95.22	75.15	**65.09**	78.83	0.00	**99.25**	49.62

Boldface indicates the best result. The proposed SRBNet achieves the best performance in five categories.

2) Performance Comparison of Each Category: To demonstrate the segmentation performance for each class, the mean Dice and the mean IoU of SRBNet and some comparison methods are shown in Table VI. The test set of EndoVis 2018 contains ten categories. The SRBNet achieves excellent performance in most categories. It takes first place in five categories. The above results demonstrate that the proposed SRBNet achieves state-of-the-art performance on EndoVis 2018.

3) Mask Visualization: Segmentation results are visualized to make the results more intuitive. The predictions of PSPNet [31], SRBNet, and ground truth are shown in Fig. 5. In images 1 and 2, the color features of the two biological tissues and the background are very similar. PSPNet cannot segment biological tissues from the background well. Compared with it, the segmentation results of the proposed SRBNet are more complete and closer to the ground truth. This proves that LBFM can effectively learn fine-grained features, thereby enhancing the feature recognition ability of the model. In images 3, 4, and 5, the specular reflection changes the color and texture of the object, which makes the objects indistinguishable. It can be found that the proposed SRBNet is less affected in the light interference area compared to PSPNet. This is because SSRM captures long-range semantic dependencies to infer the semantic features of the target from neighboring pixels. The above visualization results prove the advanced performance of the proposed method.

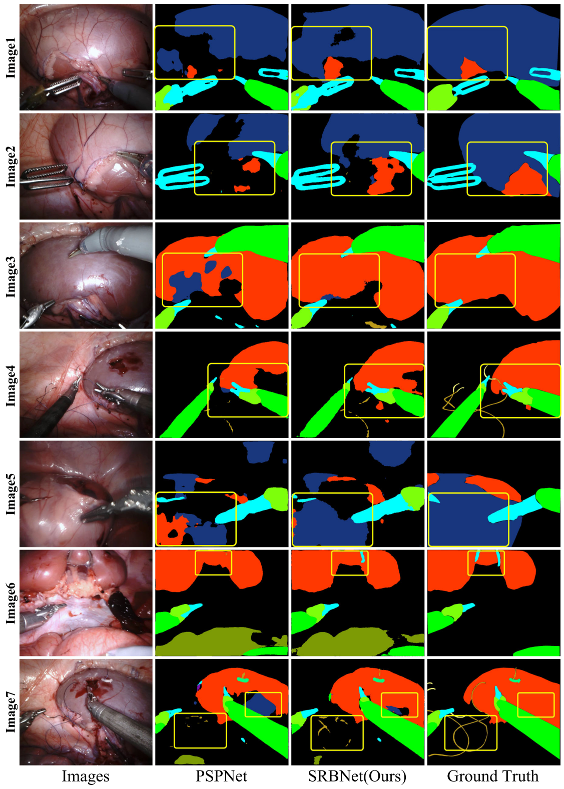


Fig. 5. Segmentation results of the proposed SRBNet. Various objects are represented by different colors. The features of biological tissues are very similar. Nevertheless, SRBNet can identify them well. Compared with PSPNet, SRBNet has more complete segmentation results and fewer recognition errors.

V. DISCUSSION AND CONCLUSION

In this paper, the SRBNet including space squeeze reasoning and low-rank bilinear feature fusion is proposed for surgical image segmentation. In it, the space squeeze reasoning module (SSRM) squeezes feature maps from different directions to encode multi-directional long-range semantic dependencies, helping to address the local similarity issue. Besides, the low-rank bilinear fusion module (LBFM) is proposed to fuse low-level and high-level features. It integrates features of different levels to boost the distinction between different semantic features. A series of experiments are set up to verify their performance. The experimental results show that SSRM can effectively solve the local similarity issue and significantly improve segmentation accuracy. Besides, LBFM has also been verified to help learn fine-grained features. The proposed SRBNet achieves state-of-the-art performance on CataSeg and EndoVis 2018.

The proposed method can be applied in computer-assisted surgery. The segmentation result can provide the doctor with visual cues during the operation, such as marking the target biological tissue and enhancing the display of small surgical instruments. The proposed method combined with the binocular vision algorithm can locate surgical instruments and biological tissues in three-dimensional space. The three-dimensional position information can be used for robot navigation.

However, the proposed SRBNet also has a limitation. The computational complexity of SRBNet is relatively high. But surgical robots often have limited computing resources. This limitation of SRBNet is not conducive to its deployment on surgical robots. Model compression methods can reduce the computational complexity of the model and maintain its accuracy. Thus, model compression is also a direction worth exploring. In the future, we will study the model compression methods based on knowledge distillation. Moreover, we will explore the application of the proposed method and binocular vision algorithm to locate surgical instruments and biological tissues for surgical robot navigation.

References

[1] I. Luengo *et al.*, "2020 cataracts semantic segmentation challenge," 2021, *arXiv:2110.10965*.

[2] M. Allan *et al.*, "2018 robotic scene segmentation challenge," 2020, *arXiv:2001.11190*.

[3] M. Allan *et al.*, "2017 robotic instrument segmentation challenge," 2019, *arXiv:1902.06426*.

[4] H. Su, Y. Hu, H. R. Karimi, A. Knoll, G. Ferrigno, and E. De Momi, "Improved recurrent neural network-based manipulator control with remote center of motion constraints: Experimental results," *Neural Netw.*, vol. 131, pp. 291–299, 2020.

[5] H. Su, A. Mariani, S. E. Ovur, A. Menciassi, G. Ferrigno, and E. De Momi, "Toward teaching by demonstration for robot-assisted minimally invasive surgery," *IEEE Trans. Automat. Sci. Eng.*, vol. 18, no. 2, pp. 484–494, Apr. 2021.

[6] Y. Xu *et al.*, "Artificial intelligence: A powerful paradigm for scientific research," *Innov.*, vol. 2, no. 4, 2021, Art. no. 100179.

[7] W. Song, W. Kang, Y. Yang, L. Fang, C. Liu, and X. Liu, "TDS-net: Towards fast dynamic random hand gesture authentication via temporal difference symbiotic neural network," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2021, pp. 1–8.

[8] M. Islam, V. Vibashan, C. M. Lim, and H. Ren, "ST-MTL: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101837.

[9] F. Qin, S. Lin, Y. Li, R. A. Bly, K. S. Moe, and B. Hannaford, "Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 6639–6646, 2020.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[11] Z.-L. Ni *et al.*, "BARNet: Bilinear attention network with adaptive receptive field for surgical instrument segmentation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 832–838.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[13] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[15] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[16] F. Qin, Y. Li, Y. Su, D. Xu, and B. Hannaford, "Surgical instrument segmentation for endoscopic vision with data fusion of CNN prediction and kinematic pose," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 9821–9827.

[17] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 440–448.

[18] Z.-L. Ni *et al.*, "RAUNet: Residual attention u-net for semantic segmentation of cataract surgical instruments," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 139–149.

[19] M. Attia, M. Hossny, S. Nahavandi, and H. Asadi, "Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2017, pp. 3373–3378.

[20] M. Islam, D. A. Atputharuban, R. Ramesh, and H. Ren, "Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 2188–2195, Apr. 2019.

[21] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.

[22] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.

[23] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *Brit. Mach. Vis. Conf.*, 2018, p. 285.

[24] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

[25] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *5th Int. Conf. Learn. Represent.*, 2017.

[26] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 574–589.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[29] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[30] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[32] E. Flouty *et al.*, "CADIS: Cataract dataset for image segmentation," *Med. Image Anal.*, vol. 71, p. 102053, 2021.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.