# BARNet: Bilinear Attention Network with Adaptive Receptive Fields for Surgical Instrument Segmentation

**Zhen-Liang Ni**[1,2] , **Gui-Bin Bian**[1,2*] , **Guan-An Wang**[1,2] , **Xiao-Hu Zhou**[1] ,
**Zeng-Guang Hou**[1,2,3] , **Xiao-Liang Xie**[1] , **Zhen Li**[1] and **Yu-Han Wang**[1]
[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{nizhenliang2017, guibin.bian, wangguanan2015, xiaohu.zhou, zengguang.hou, xiaoliang.xie, zhen.li}@ia.ac.cn, yuhanwang96318@gmail.com

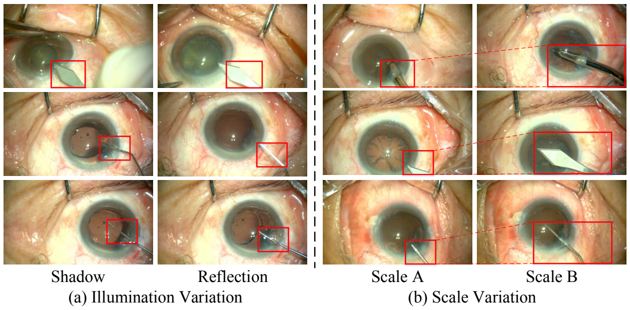
## Abstract

Surgical instrument segmentation is crucial for computer-assisted surgery. Different from common object segmentation, it is more challenging due to the large illumination variation and scale variation in the surgical scenes. In this paper, we propose a bilinear attention network with adaptive receptive fields to address these two issues. To deal with the illumination variation, the bilinear attention module models global contexts and semantic dependencies between pixels by capturing second-order statistics. With them, semantic features in challenging areas can be inferred from their neighbors, and the distinction of various semantics can be boosted. To adapt to the scale variation, our adaptive receptive field module aggregates multi-scale features and selects receptive fields adaptively. Specifically, it models the semantic relationships between channels to choose feature maps with appropriate scales, changing the receptive field of subsequent convolutions. The proposed network achieves the best performance 97.47% mean IoU on Cata7. It also takes the first place on EndoVis 2017, exceeding the second place by 10.10% mean IoU.

## 1 Introduction

In recent years, there has been significant progress in minimally invasive robotic surgery and computer-assisted microsurgery. Semantic segmentation of surgical instrument plays a crucial role in assisted surgery. It can accurately locate the surgical instrument and estimate its pose, which is essential for surgical robot control [Allan *et al.*, 2019]. Furthermore, the mask generated by semantic segmentation offers numerous solutions to assist surgery, such as real-time surgical reminder, objective assessment of surgical skills, surgical report generation and surgical workflow optimization [Sarikaya *et al.*, 2017]. These applications can improve the safety of surgery and reduce the workload of doctors.

Compared with common object segmentation, the complex surgical scenes make accurate instrument segmentation more challenging. The first difficulty is the large illumination variation caused by different light angles and occlusions. As shown in Figure 1 (a), surgical instruments tend to be whiter under specular reflection while the shadow makes instruments and background black. These problems seriously affect the visual representation of surgical instruments such as color and texture, impeding identifying the instruments stably. The second difficulty is the large scale variation caused by continuous movement and view changes. It leads to different shapes and scales of the same instrument. For example, the incision knife is in the shape of a triangle when the scale is small and in the shape of a polygon when the scale is large in Figure 1 (b). This issue makes the instrument segmentation more challenging.



(a) Illumination Variation
(b) Scale Variation

Figure 1: Difficulties in semantic segmentation for surgical instruments. Illumination variation changes the color and texture of instruments. Scale variation changes the size and shape of instruments.

Recently, a series of methods have been proposed for the semantic segmentation of surgical instruments. 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] introduced Recurrent Neural Network to capture global contexts and expand the receptive field. MF-TAPNet [Jin *et al.*, 2019] adopted optical flow as temporal prior to provide a reliable indication of the instrument location and shape for accurate segmentation. ToolNet-C combined with the kinematic pose information to get the accurate silhouette mask [Qin *et al.*, 2019]. However, most of these methods focus on expanding the receptive field and capturing shape prior while fail to address the illumina-

[*] Corresponding Author

tion variation and scale variation issues.

To address the issues mentioned above, we reconsider the features affected by them. Illumination variation affects the color and texture appearance, making identifying instruments harder. Considering that a surgical instrument is spatially continuous, we can infer the target region according to its neighbor pixels based on semantic dependencies and global contexts. To this end, a bilinear attention module (BAM) is proposed, which is based on bilinear pooling to model semantic dependencies and aggregate global contexts. Bilinear pooling can capture second-order statistics to encode complex semantic dependencies, helping to improve feature representations. Furthermore, attention features generated by bilinear pooling are adaptively distributed to each location, making every pixel feel global contexts. In this way, semantic features in reflective or shaded areas can be inferred based on semantic dependencies and global contexts, dealing with the illumination variation.

Besides, the scale variation changes the shape and size of surgical instruments. Thus, we propose an adaptive receptive field module (ARF) to select and fuse feature maps with different scales adaptively. By doing so, we can cover various scales and make predictions more reliable. Specifically, ARF includes two branches. The former learns semantic relationships among channels, and the latter aggregates multi-scale features. Channel-wise semantic relationships are applied to select feature maps with appropriate sizes. Since kernels with the same size have different receptive fields on feature maps with various sizes, this module can select appropriate receptive fields for instruments at various scales by selecting specific feature maps, adapting to the scale variation. Moreover, dense connections across scales are introduced to propagate multi-scale features, which can cover a larger scale range.

Based on the above analysis, the bilinear attention network with adaptive receptive fields, named BARNet, is proposed. The contributions of this work are as follows:

- We propose the bilinear attention module to model semantic dependencies and aggregate global contexts for inferring the semantic features in challenging regions.
- We design the adaptive receptive field module to select the appropriate receptive field adaptively, adapting to the scale variation of instruments.
- The proposed network achieves state-of-the-art performance 97.47% mean IoU on Cata7 and takes the first place on EndoVis 2017, exceeding the second place by 10.10% mean IoU.

## 2 Related Work

### 2.1 Attention

Attention mechanisms are widely used in semantic segmentation tasks. Some attention models extracted attention features based on first-order operations such as global average pooling and convolution pooling. SENet [Hu *et al.*, 2018] applied global average pooling to capture global contexts and model semantic relationships between channels. AGRNet [Zhang *et al.*, 2018] utilized convolution pooling to generate attention features. Besides, some works applied second-order models to encode complex semantic dependencies. For example, A2Net [Chen *et al.*, 2018] was based on bilinear pooling to capture second-order statistics and model semantic dependencies. The bilinear attention networks [Kim *et al.*, 2018] learned bilinear attention distributions by utilizing the low-rank bilinear pooling technique. These methods suggest that bilinear models can capture abundant semantic relationships to improve feature representation. Different from the above methods, we design a novel encoder-decoder architecture to aggregate and distribute attention features, which can achieve better results.

### 2.2 Adaptive Receptive Field and Pyramid Features

Pyramid features play a critical role in the segmentation of multi-scale targets. A range of methods improved feature representation by aggregating multi-scale features. For example, PSPNet [Zhao *et al.*, 2017] adopted pyramid pooling to extract multi-scale features. DeepLabV2 [Chen *et al.*, 2017] made use of dilated convolutions with different dilation rates to achieve multi-scale features. Feature pyramid network [Lin *et al.*, 2017b] laterally propagated multi-scale features for building feature maps with rich semantic information at all scales. However, these methods only concatenate multi-scale features together while they fail to select appropriate scales for a specific task. SKNet utilized different size convolution kernels to generate feature maps with different receptive fields and filter them [Li *et al.*, 2019], which provided a solution for the adaptive selection of receptive fields. Different from the above methods, we directly select multi-scale features instead of different size kernels and do not need to generate new features, reducing calculation costs. Besides, multi-scale features are generated by dense connections across scales, which can cover more scale ranges and boost the reuse of multi-scale features [Yang *et al.*, 2018].

## 3 BARNet

### 3.1 Overview

Illumination variation leads to the change in color and texture of instruments. Scale variation varies the size and shape of instruments. To address these issues, a bilinear attention network with adaptive receptive fields is proposed, which can capture semantic dependencies for inferring semantic features in challenging areas and select receptive fields adaptively for adapting to the scale variation.

The overall network architecture is shown in Figure 2. The bilinear attention module models semantic dependencies and generates attention features. The adaptive receptive field module aggregates multi-scale features and selects appropriate receptive field for instruments with different scales. Besides, cross-scale dense connections are introduced to propagate multi-scale features, which can also improve information flow and boost the reuse of features. Since the input of the adaptive receptive field module should be multi-scale features, we do not use it when the feature map is small. Residual Network pre-trained on the ImageNet is adopted as the backbone network.

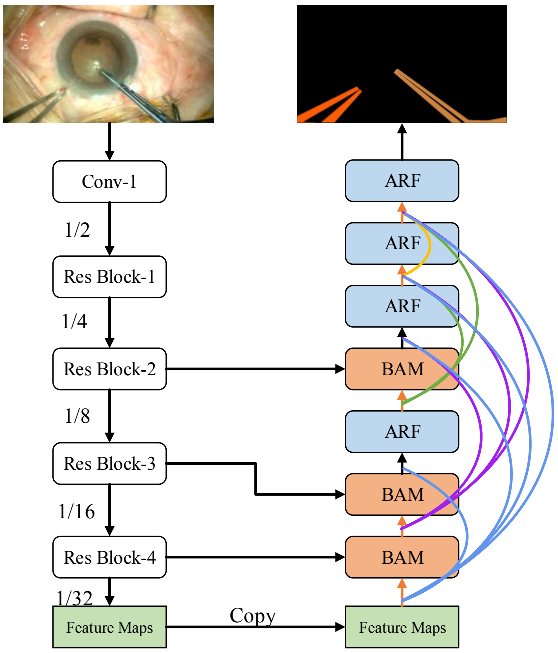


Figure 2: The architecture of BARNet. It contains two critical modules: the bilinear attention module and the adaptive receptive field module. Dense connections across scales are introduced to propagate multi-scale features.

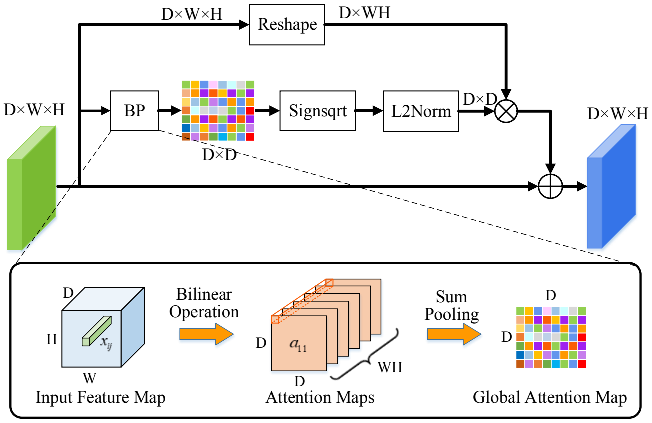


Figure 3: The architecture of the bilinear attention module. ⊗ denotes matrix multiplication. ⊕ denotes addition.

## 3.2 Bilinear Attention Module

Illumination variation changes the color and texture of surgical instruments. The network cannot directly utilize these features to identify surgical instruments, making it difficult to segment them. To solve this issue, the bilinear attention module is proposed to model semantic dependencies and capture global contexts. It is based on bilinear pooling to capture second-order statistics [Lin and Maji, 2017], which helps to boost the distinction between different semantics. Thus, the bilinear attention module can encode more complex dependencies. Besides, a decoder is designed to distribute global contexts to each location adaptively. The bilinear attention module is shown in Figure 3, including three parts: encoding, normalization, and decoding.

$$Z = F_{decode}\left(F_{norm}(F_{bp}(X))\right) \tag{1}$$

The first step is to model semantic dependencies and capture global contexts. Bilinear pooling is utilized to achieve this goal, which is shown in Eq.(2). It calculates the outer product of pixel feature vector pair $(x_{ij}, y_{ij})$ to capture second-order statistics and generate attention map $a_{ij}$, where $x_{ij}, y_{ij} \in R^{D\times 1}$ and $a_{ij} \in R^{D\times D}$. Each attention map represents the features of a pixel. These attention maps are concatenated together to encode spatial semantic dependencies. Then, sum pooling is performed to generate the global attention map $A \in R^{D\times D}$. Semantic features in all locations are encoded into each element of the global attention map, making each element feel global information. In this way, the bilinear attention module models semantic dependencies and aggregates global contexts.

$$A = F_{bp}(X,Y) = XY^T = \sum_{i=1}^{W}\sum_{j=1}^{H} x_{ij}{y_{ij}}^T \tag{2}$$

where $X, Y \in R^{D\times W\times H}$. In this work, we set $X = Y$.

Then, the global attention map $A$ is normalized to further improve its feature representation. The element-wise signed square-root and $\ell_2$ normalization are performed to normalize it, which can improve performance in practice. Also, these operations are piecewise differentiable, which can be applied to end-to-end training [Lin and Maji, 2017].

$$A' = F_{norm}(A) = sign(A)\sqrt{|A|}\Big/\left\|sign(A)\sqrt{|A|}\right\|_2 \tag{3}$$

The last step is to distribute global attention features to each pixel of the input feature map and make semantic features of each pixel calibrated by them. The input feature map $X$ is reshaped into $\overline{X} \in R^{D\times WH}$. As shown in Eq.(4), the global attention map $A'$ are distributed to each location of the input feature map by multiplication with $\overline{X}$. Besides, $X$ is added to further adjust semantics.

$$Z = F_{decode}(A, X) = A' \times \overline{X} + X \tag{4}$$

The bilinear attention module allows each pixel of feature maps to feel global contexts. Semantic features in reflective or shaded areas can be inferred based on global contexts and semantic dependencies, dealing with the illumination variation issue. It encodes semantic dependencies in the form of second-order statistics, which boosts the distinction between various semantics and improves feature representation. Furthermore, this module only performs matrix operations and does not include convolution operations. Thus, it does not add any parameters and can be easily inserted into other networks.

## 3.3 Adaptive Receptive Field Module

Since the surgical instruments are continually moving during the surgery, their shape and scale are constantly changing. Adaptive receptive fields can help the network adapt to scale

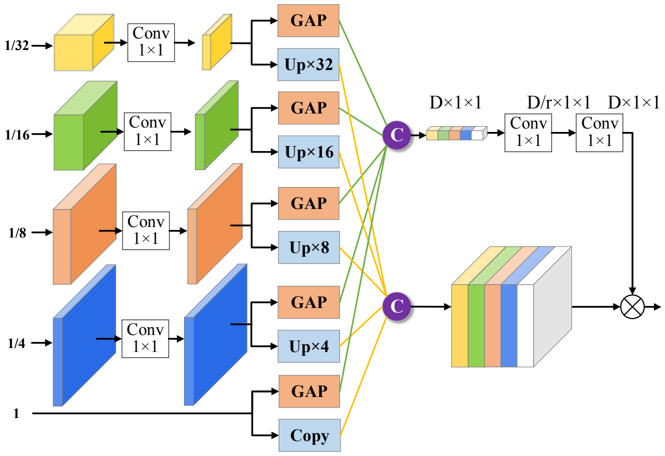


Figure 4: The architecture of the adaptive receptive field module. $\otimes$ denotes broadcast Hadamard product.

variation and learn more detailed features. Thus, the adaptive receptive field module is proposed to aggregate multi-scale features and select the appropriate receptive field for instruments with various scales. The global average pooling is introduced to model the semantic relationships between channels and generate the weight vector. The weight vector can highlight feature maps which have an appropriate scale. Since kernels with the same size have different receptive fields on feature maps at different scales, the receptive field of subsequent convolutions can be determined by selecting feature maps in a specific size. In this way, the adaptive receptive field module can select the receptive field adaptively according to the semantic relationships between channels.

The adaptive receptive field module is illustrated in Figure 4. Take the input feature maps with five scales as an example. First, $1\times1$ convolution is performed on low-scale feature maps to adjust the channel dimension to $N$, which helps to integrate multi-scale features and reduce computational costs. The maximum scale feature map does not compress the channels to preserve high-level semantic features as much as possible. $N$ can be selected according to the complexity of the network. In this paper, we set $N$ to 8. Then, multi-scale features are fed into two branches, one of which models the semantic relationships between channels and another one aggregates multi-scale features.

Specifically, in the first branch, multi-scale features are fed into global average pooling to generate vectors that encode semantic relationships [Hu *et al.*, 2018]. These vectors are concatenated together and go through two convolution layers to further extract semantic relationships. In this way, we obtain the weight vector, which represents the degree of semantic responses for different feature maps. In the second branch, multi-scale features are aggregated by upsampling and concatenation, generating the pyramid feature. The pyramid feature is multiplied by the weight vector to select the feature map with a larger response. In this way, it can adjust the receptive field of subsequent convolutions adaptively.

$$P_S = S \otimes P \tag{5}$$

where $\otimes$ refers to broadcast Hadamard product, $P$ represents the pyramid feature and $S$ denotes the weight vector.

$$P = H\left([x_k, f_{2^k}(x_0), f_{2^{k-1}}(x_1), \ldots, f_{2^2}(x_{k-2})]\right) \tag{6}$$

where $x_k$ represents the $k$-th feature map. $f_{2^k}$ denotes the $1\times1$ convolution and the $2^k\times$ upsampling. $H$ refers to concatenation.

## 4 Experiments

### 4.1 Dataset

The cataract surgical instrument dataset, Cata7, is used to evaluate our network. This dataset contains seven cataract surgery videos. To reduce redundancy, each video is downsampled from 30 fps to 1 fps. The resolution of the image is 1920×1080 pixels. It contains 2500 images, 1800 of which are used for training, and the others are used for testing. There are ten cataract surgical instruments in this dataset.

EndoVis 2017 dataset [Allan *et al.*, 2019] is from the 2017 MICCAI EndoVis Robotic Instrument Segmentation Challenge, which is based on endoscopic surgery. It contains 3000 images with a resolution of 1280×1024, which contains 1800 images for training and 1200 images for the test. There are seven types of surgical instruments in EndoVis 2017.

### 4.2 Implementation Details

All experiments are implemented on two Nvidia Titan X. The ResNet34 pre-trained by ImageNet is used as the encoder. Adam, with batch size eight, is used to train our network. The learning rate is dynamically adjusted during training. The initial learning rate on the Cata7 is $8 \times 10^{-6}$ and the initial learning rate on the EndoVis 2017 is $5 \times 10^{-5}$. For every 30 iterations, the learning rate is multiplied by 0.8. Due to limited computing resources, each image in the Cata7 is resized to 960×544, and images in the EndoVis 2017 are resized to 640×512. Data augmentation is only performed on the Cata7. The selected samples are randomly rotated, shifted, and flipped. 800 images are obtained for train by data augmentation. To objectively evaluate our model, Dice and Intersection-over-Union (IoU) are selected as the evaluation metric.

To address the class imbalance issue, we use a hybrid loss consisting of cross-entropy and Dice [Ni *et al.*, 2019], which is shown in Eq.(7).

$$Loss = (1-\alpha)H - \alpha \ln(D), \alpha \in [0, 1] \tag{7}$$

where $H$ refers to cross-entropy and $D$ denotes Dice loss. $\alpha$ is a weight used to balance cross-entropy and Dice loss. It is set to 0.2 for best training results.

### 4.3 Ablation Study Based on Cata7

**Ablation Study for Bilinear Attention Module**

Bilinear attention module (BAM) is introduced to capture the second-order statistics and model long-range semantic dependencies. To verify its performance, some experiments are performed, as shown in Table 1.

BARNet without BAM and ARF is used as the basic network. Compared with the basic network, the network using BAM achieves an increase of 4.66% mean IoU and 2.69%

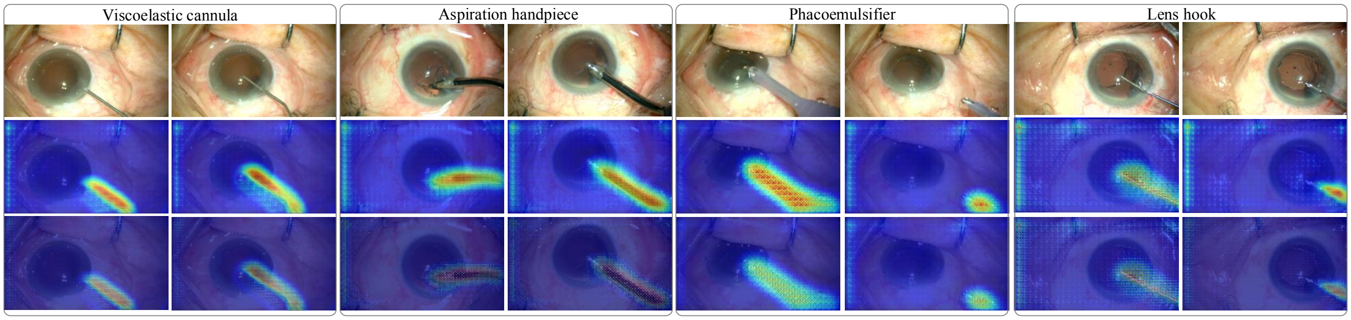


Figure 5: Visualization of attention feature maps. From top to bottom: original image, attention feature maps generated by bilinear attention module, input feature maps of bilinear attention module. Compared with input feature maps, attention feature maps highlight areas where the instruments are located more, which verifies the validity of the bilinear attention module.

| Method | ARF | BAM | mDice(%) | mIoU(%) | Param. |
|---|---|---|---|---|---|
| Basic | | | 95.12 | 91.31 | 21.80M |
| Basic | | ✓ | 97.81 | 95.97 | 21.80M |
| Basic | ✓ | | 98.06 | 96.28 | 21.90M |
| Basic | ✓ | ✓ | 98.68 | 97.47 | 21.90M |

Table 1: Ablation experiments for the bilinear attention module and the adaptive receptive field Module on Cata7.

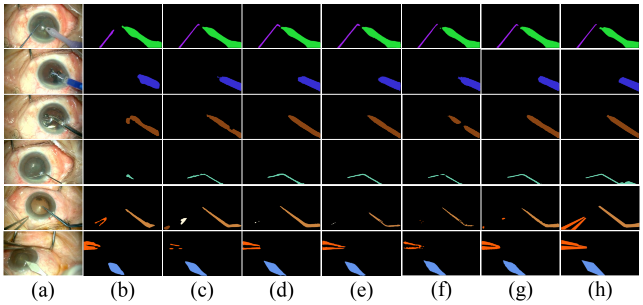


Figure 6: Visualization of segmentation results for different methods. (a) Image; (b) U-Net; (c) TernausNet; (d) LinkNet; (e) Without BAM; (f) Without ARF; (g) BARNet (Ours); (h) Groud truth.

mean Dice. When using ARF, employing BAM brings a 1.19% increase on mean IoU and 0.62% increase on mean Dice. Besides, BAM does not add any parameters since it does not contain convolution operations, as shown in Table 1. These experiments demonstrate that BAM can significantly improve network performance without adding any parameters.

To further prove the validity of the bilinear attention module, we visualize feature maps of its inputs and outputs in Figure 5. Compared with input feature maps, output feature maps of the bilinear attention module highlight the regions containing instruments more, proving that BAM effectively models semantic dependencies and improves feature representation. Also, we visualize the segmentation results of the network without BAM, which is shown in Figure 6 (e). There is incomplete segmentation in these results, and part of the surgical instrument is identified as background. The network employing BAM achieves excellent results, whose masks are relatively complete and the same as the ground truth.

| Method | mDice | mIoU | Param. |
|---|---|---|---|
| U-Net [Ronneberger *et al.*, 2015] | 86.83 | 78.21 | 7.85M |
| RefineNet [Lin *et al.*, 2017a] | 93.53 | 88.41 | 25.75M |
| LinkNet [Chaurasia and Culurciello, 2017] | 94.63 | 91.31 | 21.80M |
| TernausNet [Iglovikov and Shvets, 2018] | 96.40 | 92.98 | 25.36M |
| RAUNet [Ni *et al.*, 2019] | 97.71 | 95.62 | 22.06M |
| BARNet(Ours) | 98.68 | 97.47 | 21.90M |

Table 2: Segmentation results of different methods on Cata7.

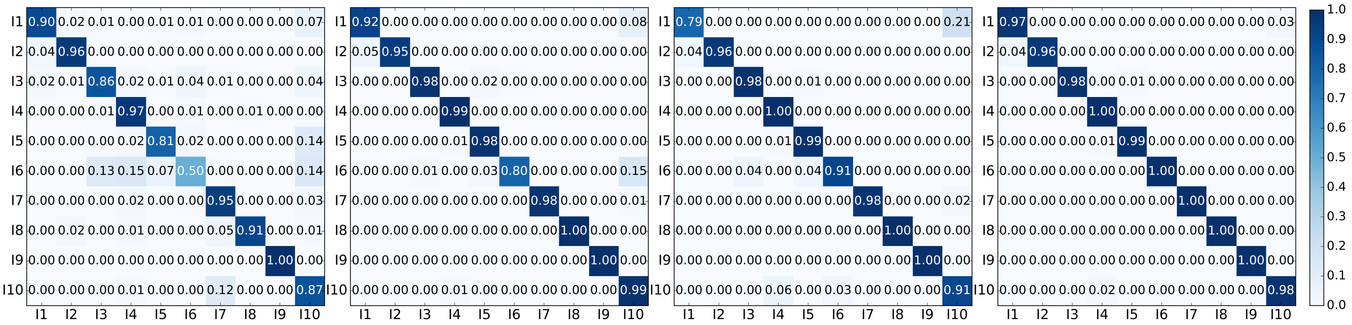
### Ablation Study for Adaptive Receptive Field Module

Adaptive Receptive Field Module (ARF) is designed to adaptively select the receptive field, adapting to the scale variation of instruments. A series of experiments are set up to verify its performance. As shown in Table 1, the network using ARF achieves an increase of 4.97% mean IoU and 2.94% mean Dice compared to the basic network. When using BAM, employing ARF brings a 1.50% increase on mean IoU and 0.87% increase on mean Dice. Furthermore, ARF only adds 0.1M parameters which only account for 0.46% of the basic network. To give a more intuitive result, we visualize some results of the network without ARF, which is shown in Figure 6 (f). The network without ARF has poorer segmentation performance than BARNet, indicating the effectiveness of the ARF. The above results suggest that ARF can significantly improve segmentation accuracy with very few parameters.

## 4.4 Comparison with State-of-the-art on Cata7

A series of comparative experiments are performed to evaluate the performance of BARNet. BARNet achieves state-of-the-art performance 97.47% mean IoU and 98.68% mean Dice, exceeding the second-ranking method by 1.85% on mean IoU and 0.97% on mean Dice. The performance of other methods is much poorer than BARNet, which demonstrates its excellent performance.

To further evaluate the segmentation performance of the proposed method for each type of surgical instrument, the confusion matrix based on pixel classification is shown in Figure 7. We find that our method achieves excellent performance on every type of instrument. Especially, the proposed method outperforms other methods by a significant margin on primary incision knife (I1), lens hook (I6) and bonn forceps (I10). The surface of these surgical instruments is prone



(a) U-Net (b) TernausNet (c) LinkNet (d) BARNet(Ours)

Figure 7: The confusion matrix based on pixel classification. The numbers on the diagonal are the pixel accuracy of each class. The X and Y-axis represent prediction and ground truth, respectively.

| | 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 | 0.104 | **0.801** | 0.919 | **0.934** | **0.830** | **0.615** | **0.534** | **0.897** | 0.352 | **0.810** | **0.643** |

Table 3: Segmentation results on Endovis 2017 dataset. BARNet achieves 64.30% mean IoU and takes the first place. NCT, UB and UA are the university abbreviation of the participating team.

to specular reflections due to their special material, making it more difficult to segment them. The bilinear attention module can model complex semantic dependencies and infer the semantic features in reflection and shadow regions, addressing the illumination variation issue. Thus, our network achieves better performance on these three instruments.

### 4.5 The Results on EndoVis 2017

To further verify the performance of BARNet, it is evaluated on the Endovis 2017. The test set consists of 10 video sequences. Datasets 1-8 contain 75 images, respectively. Both datasets 9 and 10 contain 300 images. The test results are reported in Table 3. TernausNet [Iglovikov and Shvets, 2018], ToolNet [García-Peraza-Herrera *et al.*, 2017], SegNet [Badrinarayanan *et al.*, 2017] and three other methods are tested. All test results are from the MICCAI EndoVis challenge 2017 [Allan *et al.*, 2019].

BARNet achieves 64.30% mean IoU, which outperforms existing methods. The second-ranking method, TernausNet, achieves 54.20% mean IoU. Compare with this method, our network achieves 10.10% gain on mean IoU. The performance of other methods is much poorer than BARNet. Furthermore, BARNet achieves the best results in seven video sequences and takes the second place in two video sequences. These results show that BARNet achieves state-of-the-art performance on this dataset. To give intuitive results, the segmentation results of BARNet are visualized in Figure 8. We find multiple specular reflections and shadows in the figure. Besides, the scale and shape of surgical instruments are significantly varied. Despite these challenges, BARNet can still accurately segment surgical instruments, whose segmentation results are basically the same with the ground truth.

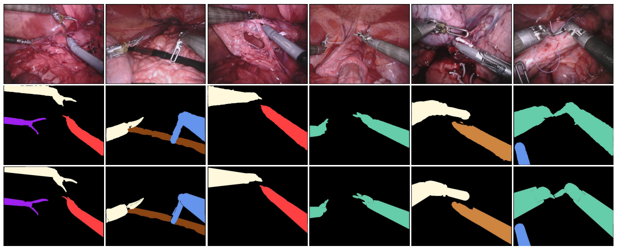


Figure 8: Visualization for segmentation results of BARNet. From top to bottom: original images, segmentation results, ground truth.

## 5 Conclusion

In this paper, we propose the BARNet for surgical instrument segmentation. In this network, the bilinear attention module captures global contexts and semantic dependencies to improve feature representation. The adaptive receptive field module selects feature maps with specific sizes to choose appropriate receptive fields. A series of ablation experiments prove that they contribute to improving network performance. Furthermore, BARNet achieves state-of-the-art performance on both 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 *et al.*, 2019] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[Attia *et al.*, 2017] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder. In *2017 IEEE International Conference on Systems, Man, and Cybernetics*, pages 3373–3378, 2017.

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[Chaurasia and Culurciello, 2017] Abhishek Chaurasia and Eugenio Culurciello. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *IEEE Visual Communications and Image Processing*, pages 1–4, 2017.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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, 2017.

[Chen *et al.*, 2018] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A^2-Nets: Double attention networks. In *Advances in Neural Information Processing Systems 31*, pages 352–361. 2018.

[García-Peraza-Herrera *et al.*, 2017] Luis C García-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, et al. Toolnet: Holistically-nested real-time segmentation of robotic surgical tools. In *the International Conference on Intelligent Robots and Systems*, pages 5717–5722, 2017.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Iglovikov and Shvets, 2018] Vladimir Iglovikov and Alexey Shvets. TernausNet: U-Net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.

[Jin *et al.*, 2019] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In *Medical Image Computing and Computer Assisted Intervention*, pages 440–448, 2019.

[Kim *et al.*, 2018] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[Li *et al.*, 2019] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 510–519, 2019.

[Lin and Maji, 2017] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. In *the British Machine Vision Conference*, pages 117.1–117.12, 2017.

[Lin *et al.*, 2017a] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Ni *et al.*, 2019] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, and Zhen Li. RAUNet: Residual attention U-Net for semantic segmentation of cataract surgical instruments. In *International Conference on Neural Information Processing*, pages 139–149. Springer, 2019.

[Qin *et al.*, 2019] Fangbo Qin, Yangming Li, Yun-Hsuan Su, De Xu, and Blake Hannaford. Surgical instrument segmentation for endoscopic vision with data fusion of rediction and kinematic pose. In *International Conference on Robotics and Automation*, pages 9821–9827, 2019.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[Sarikaya *et al.*, 2017] Duygu Sarikaya, Jason J Corso, and Khurshid A Guru. 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, 2017.

[Yang *et al.*, 2018] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.

[Zhang *et al.*, 2018] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.