2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Hierarchical Attentional Feature Fusion for Surgical Instrument Segmentation

Xiaowei Zhou[1,2], Yue Guo[1], Wenhao He[1,2], Haitao Song[1]

***Abstract*— Instrument segmentation is a crucial and challenging task for robot-assisted surgery operations. Recent commonly-used models extract feature maps in multiple scales and combine them via simple but inferior feature fusion strategies. In this paper, we propose a hierarchical attentional feature fusion scheme, which is efficient and compatible with encoder-decoder architectures. Specifically, to better combine feature maps between adjacent scales, we introduce dense pixel-wise relative attentions learned from the segmentation model; to resolve specific failure modes in predicted masks, we integrate the above attentional feature fusion strategy based on position-channel-aware parallel attention into the decoder. Extensive experimental results evaluated on three datasets from MICCAI 2017 EndoVis Challenge demonstrate that our model outperforms other state-of-the-art counterparts by a large margin.**

***Index Terms*— Hierarchical Attention UNet, Attentional feature fusion, Position-channel-aware parallel attention, Instrument segmentation**

I. Introduction

Recently, robot-assisted minimally invasive surgery has gradually become an active research area because of its potentials to improve the stability and safety of surgical operations [1], [2]. Instrument segmentation is a fundamental task for further interactions between the robot and the environment. However, accurate instrument segmentation from a static image remains challenging if context information can not be fully captured and understood, especially when the environment becomes complicated by unpredictable factors such as motion blurs of instruments, occlusions by blood, or auxiliary tools, and different lighting conditions [3].

Traditional instrument segmentation models usually capture global structures and local details by combining low-level and high-level feature maps with brute-force element-wise computations, which ignores their implicit relationships [4]. Even though attention becomes a popular component to discover relations among different feature maps, it is limited when applied on the same scale. To alleviate this issue, it can be used to fuse feature maps in different layers, and recent works mainly focus on the design combined with global channel attention modules [5], [6]. However, all the spatial information is compressed into one scalar, so such a module may not be suitable for describing regions of subtle but unique attributes of instruments in our task.

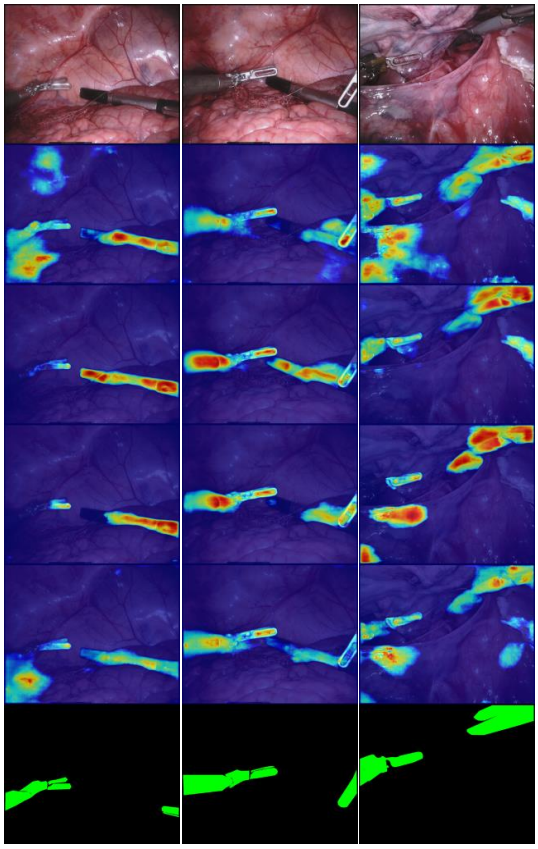


Fig. 1. Comparisons of attentional feature fusion strategies for binary segmentation: inputs, class activation maps generated by AFF-MSCAN, RAFF-MSCAN, RAFF-CBAM, and RAFF-PPA (ours), and ground-truth masks are displayed from top to bottom, respectively. It can be seen that more reasonable attentions are paid on instruments with our method.

Motivated by the attentional feature fusion [7], in this paper, we propose a hierarchical attentional feature fusion scheme, which is naturally suitable for handling feature maps across different scales in an encoder-decoder network. Specifically, position-channel-aware parallel attention is designed to solve the large gap between spatial attention and channel attention for feature maps from adjacent scales, and the hierarchical attentional feature fusion is responsible for gradually fuse maps and recover details through soft selection. As a simple model with highly-compatible modules, our Hierarchical Attention UNet, consistently outperforms other state-of-the-art networks in all the sub-tasks from MICCAI 2017 EndoVis Challenge, without bells and whistles.

*This work is supported by National Key R&D Program of China (2018YFB1306302, 2018YFB1306300, and 2018YFB1306500).

*Corresponding author: Wenhao He.

[1]Xiaowei Zhou, Yue Guo, Wenhao He, and Haitao Song are with Institute of Automation, Chinese Academy of Sciences, Beijing, China. guoyue2013@ia.ac.cn

[3]Xiaowei Zhou and Wenhao He are also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

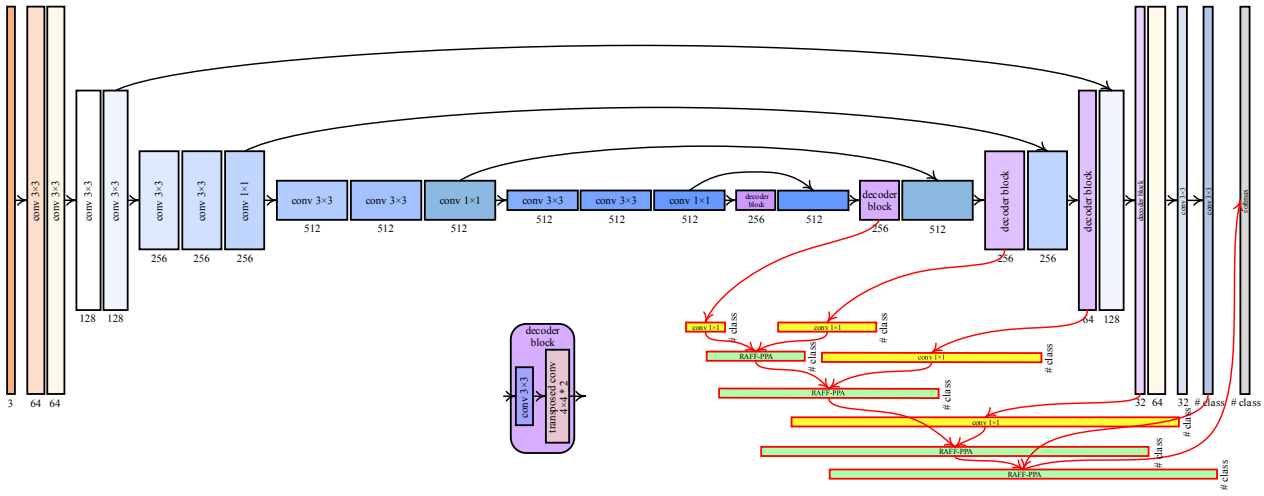


Fig. 2. Architecture of our instrument segmentation model: based on the vallina TernausNet [8], RAFF-PPAs are incorporated to fuse decoded features in adjacent scales hierarchically.

II. Related Works

A. Channel Attention Variants

Attention improves the representation power of convolutional feature maps, but more global information is preferred using the channel attention module, which mainly results from global average pooling. To alleviate this issue, channel attention can be applied with the position-sensitive spatial attention sequentially [6] or in parallel [5]. To make the channel attention module more sensitive to local information, global average pooling operations can be replaced with convolution-based residual attention [9] or non-local operations [10] in context modeling.

Self-attention is another typical module in which weight scores are computed by using weighted sum with local features, but it ignores relationships among these features. Therefore, such attentions can be learned in an interaction-aware self-attention module inspired by PCA [11], and self-attention can also be extended to multiple heads to capture different global information when spatial locations are encoded using relational position encodings [12]. Besides, temporal prior based on motion flow is integrated to segment instruments in video [13]; more generally, identical mapping is generalized to an attention module in the residual branch for image classification [14], and attentional feature fusion is capable of handling contexts for objects in various sizes [7].

B. Multi-scale Feature Selection

Feature maps in multiple independent scales may not be easily fused without attention across scales. Initially, feature maps in higher levels are combined with lower ones in UNet [15], and transposed convolution operations can be replaced with nearest upsampling ones in the decoder part of UNet [16]; features from all the intermediate layers in a fully convolutional network facilitates segmenting robotic surgical instruments [4]. However, directly pooling in feature pyramids loses pixel localizations, so object context pooling is applied to update feature maps in every scale [17], and global context prior attention is used for selecting low-level features [18]. To obtain all the attention maps hierarchically, relative masks are learned between adjacent scales [19].

In this paper, we not only incorporate multiscale attention modules in a UNet but also hierarchically fuse feature maps between adjacent scales.

III. Methods

A. Hierarchical Attention UNet

As a typical network architecture for semantic segmentation, UNet is constructed in our task, and it consists of an encoder and a decoder. Specifically, in the encoder, a 16-layer VGGNet is applied to extract convolutional feature maps [20], and there are five convolution blocks, each of which have repeated convolutional feature maps from VGGNet followed by an activation function such as ReLU. There are five deconvolution blocks in the decoder to recover feature maps from coarse levels to fine ones, each of which contains an upsampling operator followed by transposed convolution and an activation function like ReLU.

Residual attention feature fusion modules are hierarchically integrated into the decoder of an original UNet, and its architecture is illustrated in Figure 2. To demonstrate the effectiveness of the above-mentioned fusion module, no additional carefully-designed modules are added. Therefore, given an input image I, features $F_E \in \mathbb{R}^{C\times H\times W}$ from VGGNet in an intermediate layer and the corresponding features $F_D \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$ in the encoder can be respectively viewed as the local and global context.

B. Residual Attention Feature Fusion

To better fuse high-level feature maps in the decoder and low-level ones in the encoder, a specialized scheme, "Attention Feature Fusion" (AFF), is integrated to replace simple pixel-wise concatenation or summation in the vallina UNet. The mixed feature map F_M is computed using the above module as follows:

$$F_S = F_E + U(F_D)$$
$$F_M = PPA(F_S) \times F_E + (1 - PPA(F_S)) \times F_D + F_E \tag{1}$$

where F_S is the ouput of the element-wise summation after the global context is enlarged using bilinear upsampling U. $PPA(\cdot)$ denotes the output from a position-channel-aware parallel attention module to balance local and global contexts.

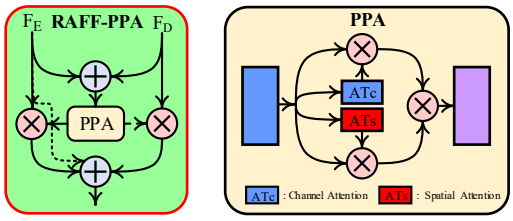


Fig. 3. Structure of hierarchical attention feature fusion: PPA outputs weights of a position-channel-aware parallel attention module, so features focused by both spatial attention and channel attention are emphasized. RAFF-PPA fuses attentional features from the low and high levels.

C. Position-channel-aware Parallel Attention

To exploit both spatial attention and channel one in the attentional feature fusion scheme, a position-channel-aware parallel attention module is constructed, as shown in Figure 3.

The local context F_E and the global one F_D are fused in our attention module:

$$PPA(F_S) = AT_S(F_S) \times AT_C(F_S) \tag{2}$$

where $AT_S(\cdot)$ and $AT_C(\cdot)$ represent spatial attention and channel attention, respectively, and they are defined as follows:

$$AT_S(F_S) = \sigma\Big(C_{1\times1}\big([P_{AVG}(F_S); P_{MAX}(F_S)]\big)\Big)$$
$$AT_C(F_S) = \sigma\Big(FFN\big(P_{AVG}(F_S)\big) + FFN\big(P_{MAX}(F_S)\big)\Big) \tag{3}$$

where $C_{1\times1}(\cdot)$ is a convolution module which kernel size is 1×1; $FFN(\cdot)$ denotes a three-layer feed forward neural network: the dimension of hidden layer is one, and those of input and output layers are the same. $P_{AVG}(\cdot)$ and $P_{MAX}(\cdot)$ represent the average pooling and max pooling operations.

D. Loss Function

To supervise training procedures of instrument segmentation models, the applied total loss function L_{total} contains a multi-category negative log likelihood loss L_{nll} and an intersection over union one $mIoU$:

$$L_{nll} = -\frac{1}{n_c}\sum_{c=1}^{n_c}\sum_{i=1}^{n_i} G_{c,i} log(P_{c,i})$$
$$L_{total} = L_{nll} - log(mIoU) \tag{4}$$

where the dataset for the current sub-task has n_c classes in total, and the predicted mask has n_i pixels. $mIoU$ is an evaluation metric that will be introduced in Section IV-C.

IV. Experiments

A. Dataset

To evaluate the performances of our model, we choose the original dataset about robotic instrument segmentation in MICCAI 2017 EndoVis Challenge [1]. Specifically, the width of a sampled image is 1920 pixels (320 pixels on both the left and right borders of the image), and its height is 1080 pixels (28 pixels respectively on the top and bottom borders). Therefore, the resolution of an input image becomes 1280×1024 after the above-mentioned black borders are cropped out.

This dataset contains a training set (1800 images) and a test set (1200 images), which has already been held out by the challenge organizers, and it has seven different kinds of instruments including six operation instruments and one robot-assisted counterpart. Based on this dataset, the segmentation tasks can be divided into three sub-tasks: binary segmentation (instrument or background), instrument segmentation (seven types of instruments), and part segmentation (shaft, clasper, and wrist).

B. Implementation Details

All the experiments are taken under the deep learning framework named Pytorch, and models are trained using four TITAN XP GPU cards. Four-fold cross-validation is used when the model training where the training set is split into 1350 images for training and 450 samples for validation.

During training, Adam [21] is selected as our optimizer, and its base learning rate is set as 0.00003. To accelerate the forward propagation speed, the model input is resized to $\frac{1}{4}$ the resolution of the image without black borders, and the batch size is eight; then they are mainly augmented using random operations including up-down and left-right flips.

During the inference, given the cropped and rescaled images, the segmentation model directly outputs the semantic mask with the same resolution, then these masks are resized to 1280×1024 for evaluation on three sub-tasks.

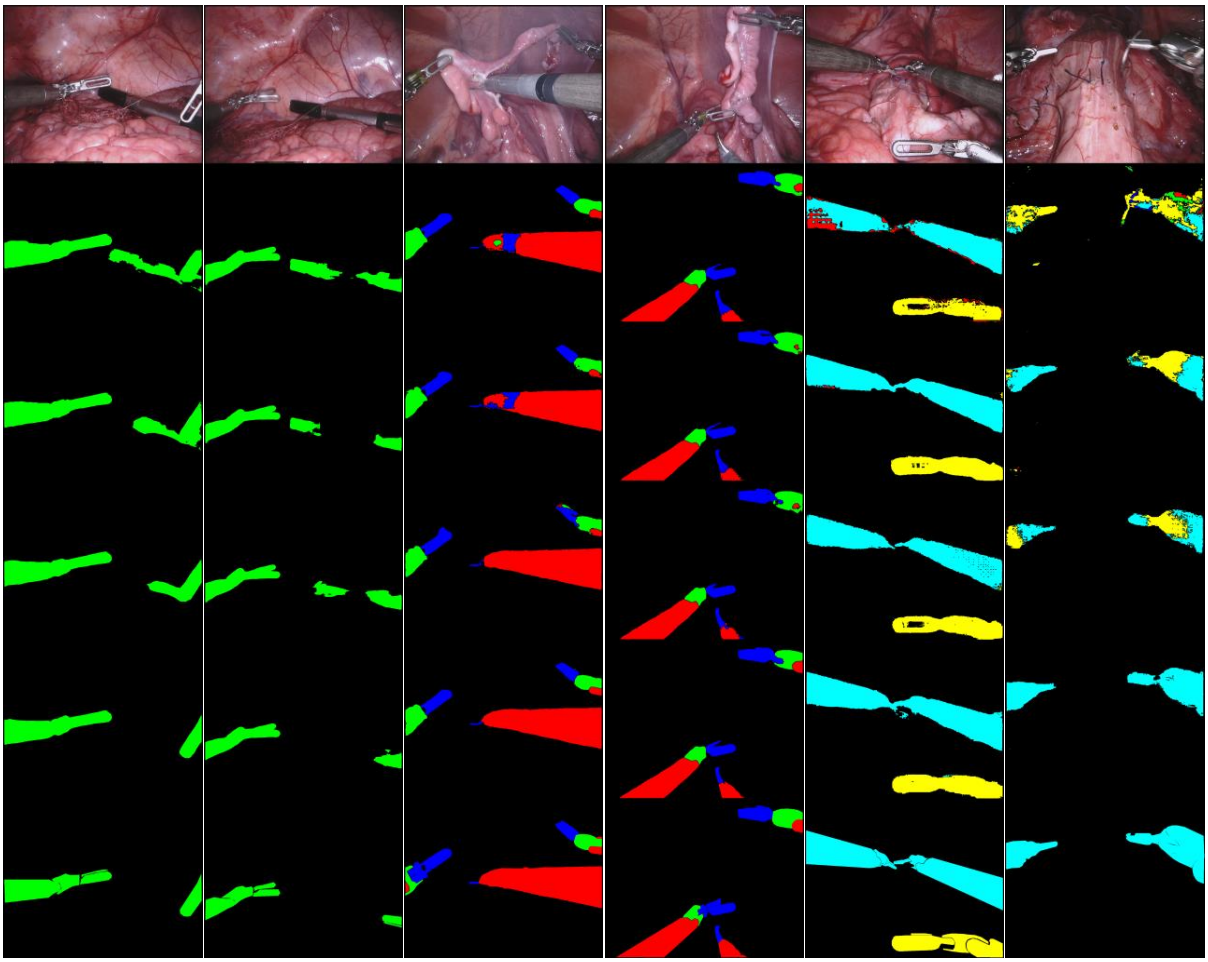
C. Metrics

Similar to metrics introduced in previous works, mIoU (mean Intersection of Union) and mDice (mean Dice Coefficient) are used for evaluation in this paper. Given a ground-truth mask G and a predicted mask P for class c, its IoU IoU_c and Dice $Dice_c$ are calculated as follows:

$$IoU_c = \frac{\sum_{i=1}^{n_i}\sum_{j=1}^{n_i} P_{c,i}G_{c,j}}{\sum_{i=1}^{n_i} P_{c,i} + \sum_{i=1}^{n_i} G_{c,i} - \sum_{i=1}^{n_i}\sum_{j=1}^{n_i} P_{c,i}G_{c,j}} \tag{5}$$

$$Dice_c = \frac{2\sum_{i=1}^{n_i}\sum_{j=1}^{n_i} P_{c,i}G_{c,j}}{\sum_{i=1}^{n_i} P_{c,i} + \sum_{i=1}^{n_i} G_{c,i}} \tag{6}$$

It is important to note that infinite small numbers to avoid zero division is ignored for simplicity. Consequently, the



(a) Binary Segmentation (b) Part Segmentation (c) Type Segmentation

Fig. 4. Results for binary, part, and type segmentations: given input images on the first row, segmentation results provided by AFF-MSCAN, RAFF-MSCAN, RAFF-CBAM, RAFF-PPA (ours), and ground-truth masks are shown from top to bottom, respectively.

average values of IoUs $mIoU$ and those of Dice Coefficents $mDice$ among all the categories are defined as follows:

$$mIoU = \frac{1}{n_c}\sum_{c=1}^{n_c} IoU_c$$
$$mDice = \frac{1}{n_c}\sum_{c=1}^{n_c} Dice_c \quad (7)$$

V. Ablation Studies

Components in our instrument segmentation model are investigated, as listed in Table I. To simplify the annotations of different modules, Attentional Feature Fusion [7], Residual Path [22], Convolutional Block Attention Module [6], Position-channel-aware Parallel Attention module are respectively abbreviated as AFF, R, CBAM, and PPA. By default, the attention block in AFF is MSCAN.

A. Effectiveness of Attentional Feature Fusion

Attentional feature fusions significantly improve binary and part segmentation performances. Specifically, mIoUs of TernausNet after adding AFF respectively increase from 83.60% to 89.73% for binary segmentation and from 65.05% to 71.02% for part segmentation. However, it slightly drops from 33.78% to 33.40% for type segmentation, while the corresponding mDice increases by 0.08%.

B. Effectiveness of Residual Path

Residual paths slightly deteriorate the binary performance but improve results of part and type segmentations. For example, compared to TernausNet with AFF-MSCAN on mIoUs, this model with RAFF-MSCAN merely drops from 89.73% to 89.65% for binary segmentation, while it slightly increases by 0.4% for part segmentation and significantly improves from 45.03% to 56.80% for type segmentation.

C. Effectiveness of Position-channel-aware Parallel Attention

Conditioned on the residual paths and attentional feature fusions, models with position-channel-aware parallel attention modules consistently improve the performances on all the three sub-tasks. For instance, TernausNet + RAFF-PPA achieves the highest mIoUs including 90.25%, 73.61%, and 53.91% for binary, part, and type segmentation sub-tasks.

TABLE I

RESULTS OF ABLATION STUDIES ON THREE SUB-TASKS (MEAN±STD).

Method	Binary Segmentation		Part Segmentation		Instrument Segmentation	
	mIoU (%)	mDice (%)	mIoU (%)	mDice (%)	mIoU (%)	mDice (%)
TernausNet [8]	83.60±15.83	90.01±12.50	65.05±17.22	75.97±16.21	33.78±**19.16**	44.95±**22.89**
TernausNet + AFF-MSCAN [7]	89.73±11.26	94.12±7.97	71.02±14.30	81.07±11.76	33.40±21.24	45.03±23.56
TernausNet + RAFF-MSCAN [7]	89.65±10.75	94.11±7.60	71.42±14.30	81.24±11.92	45.57±23.32	56.80±24.41
TernausNet + RAFF-CBAM [6]	89.79±11.35	94.14±8.11	71.94±13.93	81.73±11.85	45.39±25.11	56.28±26.34
TernausNet + RAFF-PPA	**90.25±10.39**	**94.48±7.26**	**73.61±13.57**	**83.01±11.10**	**53.91**±28.45	**63.16**±28.69

TABLE II

RESULTS OF STATE-OF-THE-ART METHODS ON THREE SUB-TASKS (MEAN±STD).

Method	Binary Segmentation		Part Segmentation		Instrument Segmentation	
	mIoU (%)	mDice (%)	mIoU (%)	mDice (%)	mIoU (%)	mDice (%)
UNet [15]	75.44±18.18	87.37±14.58	48.41±17.59	60.75±18.21	15.81±15.06	23.59±19.87
TernausNet [8]	83.60±15.83	90.01±12.50	65.05±17.22	75.97±16.21	33.78±19.16	44.95±22.89
UNetPlus [16]	83.75±15.36	90.19±11.77	65.75±16.74	76.25±15.54	34.19±**15.06**	45.32±**19.86**
MF-TAPNet [13]	87.56±16.42	93.37±12.93	67.92±16.50	77.05±16.17	36.62±22.78	48.01±25.64
Ours	**90.25±10.39**	**94.48±7.26**	**73.61±13.57**	**83.01±11.10**	**53.91**±28.45	**63.16**±28.69

Even though the standard deviation for type segmentation is relatively larger, which means on some occasions, results become better by a large margin.

VI. COMPARISON WITH STATE-OF-THE-ART METHODS

Over the baseline TernausNet, our method provides large boosts on mIoUs including 6.65%, 8.56%, and 20.13% for binary, part, and type segmentation, respectively. Experimental results in Table II also demonstrate the superiority of our method compared to other state-of-the-art ones.

VII. CONCLUSION

In this paper, we propose a hierarchical attentional feature fusion scheme that can better fuse feature maps for instrument segmentation in robotic-assisted surgery. Specifically, the residual attentional feature fusion module is highly compatible with a decoder that refines feature maps from coarse level to fine, and the position-channel-aware parallel attention that more efficiently exploits both spatial and channel attentions can be treated as a common attention module for maps between two different scales. Extensive results and ablation studies show the effectiveness of the proposed attentional feature fusion scheme and the superiority of our model.

REFERENCES

[1] M. Allan, A. Shvets, T. Kurmann, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint: 1902.06426*, 2019.

[2] M. Allan, S. Kondo, S. Bodenstedt, et al. 2018 robotic instrument segmentation challenge. *arXiv preprint: 2001.11190*, 2020.

[3] Z. Ni, G. Bian, G. Wang, et al. Pyramid attention aggregation network for semantic segmentation of surgical instruments. *Assoc. Adv. Artif. Intell.*, 2020.

[4] L. C. García-Peraza-Herrera, W. Li, L. Fidon, et al. Toolnet: Holistically-nested real-time segmentation of robotic surgical tools. *IEEE Int. Conf. Intell. Robot. Syst.*, pages 5717–5722, 2017.

[5] J. Fu, J. Liu, H. Tian, et al. Dual attention network for scene segmentation. *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3146–3154, 2019.

[6] S. Woo, J. Park, J. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. *Eur. Conf. Comput. Vis.*, pages 1–17, 2018.

[7] Y. Dai, F. Gieseke, S. Oehmcke, et al. Attentional feature fusion. *arXiv preprint: 2009.14082*, 2020.

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

[9] Y. Cao, J. Xu, S. Lin, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *IEEE Int. Conf. Comput. Vis.*, pages 1971–1980, 2019.

[10] X. Wang, R. Girshick, A. Gupta, et al. Non-local neural networks. *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7794–7803, 2018.

[11] Y. Du, C. Yuan, B. Li, et al. Interaction-aware spatio-temporal pyramid attention networks for action classification. *Eur. Conf. Comput. Vis.*, pages 388–404, 2018.

[12] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. *Int. Conf. Neural Info. Process. Syst.*, pages 5998–6008, 2017.

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

[14] F. Wang, M. Jiang, C. Qian, et al. Residual attention network for image classification. *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3156–3164, 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Med. Image Comput. Comput. Assist. Interv.*, pages 234–241, 2015.

[16] S. M. Kamrul Hasan and C. A. Linte. U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instrument. *Annu. Int. Conf. IEEE Eng. Med. Bio. Soc.*, pages 7205–7211, 2019.

[17] Y. Yuan and J. Wang. Ocnet: Object context network for scene parsing. *arXiv preprint: 1809.00916*, 2018.

[18] H. Li, P. Xiong, J. An, et al. Pyramid attention network for semantic segmentation. *arXiv preprint: 1805.10180*, 2018.

[19] A. Tao, K. Sapra, and B. Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint: 2005.10821*, 2020.

[20] A. Zisserman K. Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint: 1409.1556*, 2014.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint: 1412.6980*, 2014.

[22] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition. *arXiv preprint:1512.03385*, 2015.