RCANET: ROW-COLUMN ATTENTION NETWORK FOR SEMANTIC SEGMENTATION

Bingxu Lu[1], *Qinghua Hu*[1], *Yu Wang*[1], *Guosheng Hu*[2]

[1]Tianjin Key Lab of Machine Learning, College of Intelligence and Computing, Tianjin University, China
[2]AnyVision

ABSTRACT

Establishing high-order interactions among pixels and object parts is one of the most fundamental problems in semantic segmentation. The recent proposals are based on non-local methods which utilize the self-attention mechanism to capture the long-range correlations. However, non-local methods could be very expensive, both theoretically and experimentally. Moreover, non-local methods are typically designed to address spatial correlations rather than feature correlations across channels. In this work, we propose a Row-Column Attention Network (RCANet) to encode globally contextual information. It consists of a row-wise intra-channel attention module and a column-wise intra-channel attention module, followed by a cross-channel interaction module. We conduct experiments on two datasets: Cityscapes and ADE20K. The results show that our method is comparable to the state-of-the-art methods for semantic segmentation.

Index Terms— semantic segmentation, self-attention, contextual information

1. INTRODUCTION

Semantic segmentation aims at recognizing diverse objects and background of input images in the pixel level. It is a fundamental computer vision task applied in numerous fields such as autonomous driving and medical image analysis. As convolutional Neural Networks (CNN) have made milestone progress in object recognition task, CNN is naturally applied to semantic segmentation. For example, Fully Convolutional Network (FCN) [1] and the DeepLab family [2, 3, 4, 5] use mainstream CNNs as the feature encoder to generate pixel-wise predictions for semantic segmentation. EADNet [6] consists of multiple developed asymmetric convolution branches with different dilation rates to capture the variable shapes and scales information of an image.

To capture the long-range dependencies between pixels in the input, the Non-local Network [7] proposes a Non-local module (NL) using self-attention mechanism. The NL Network was firstly proposed to solve video recognition tasks, but it has been re-purposed to build semantic segmentation models. For example, DANet [8] develops the dual-path attention module to capture the contextual information over the channel and spatial dimensions separately; Axial-DeepLab [9] uses position-sensitive axial-attention layer for image classification and dense prediction; CCNet [10] proposes the continuous criss-cross attention module with sparse attention maps to reduce the computational complexity for high-resolution images.

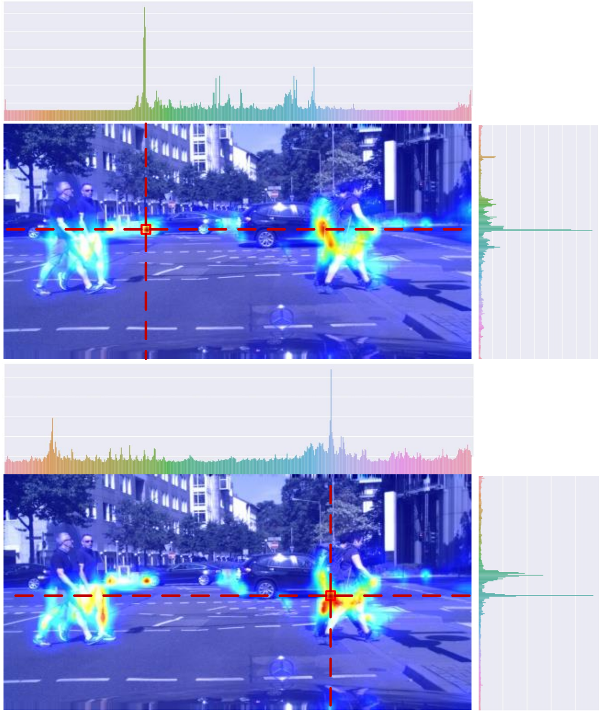

Fig. 1: The visual analysis of the attention map generated by Non-local Module and our RCANet on Cityscapes validation set.

Although the NL-based models can capture the spatial correlations, they do not model the correlations between channels. This can be an issue when we apply the NL module on imagery data as first observed by [11]. Specifically, they find that the attention maps with respect to different query locations are surprisingly similar, which suggests that the global context learned by the NL module may be independent of query position. We confirm this phenomenon by reproducing the visual experiments of [11]. Two examples of the results are shown in Fig. 1, where the two heatmaps represent the attention maps with respect to two particular pixels located on a person and a car respectively. However, their corresponding attention maps are almost identical, confirming that the NL module may not be able to capture meaningful contextual information for different query locations.

To reduce the computational cost and make attentions location-dependent, we propose a novel architecture called Row-Column Attention Network (RCANet) which contains Row-Column Atten-

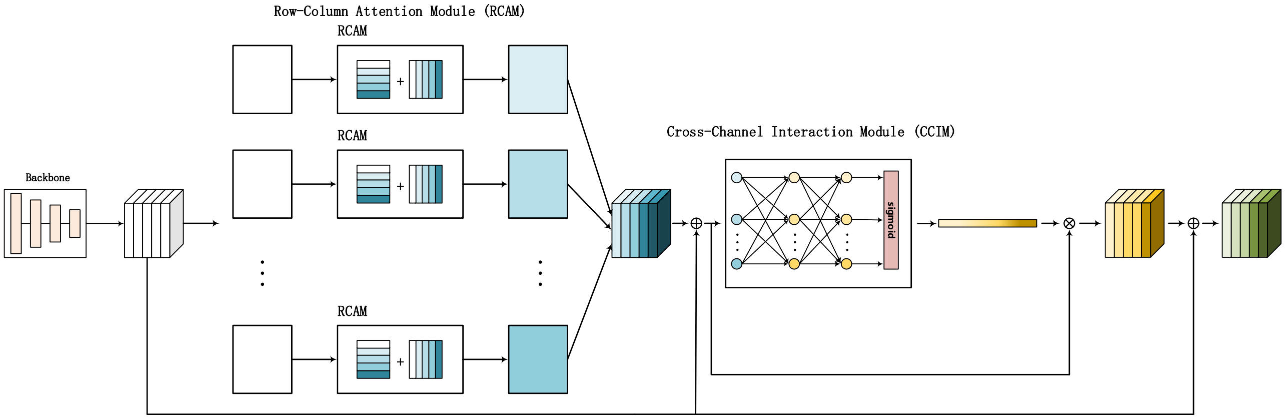
Row-Column Attention Module (RCAM)
RCAM
+
RCAM
+
RCAM
+
Backbone
Cross-Channel Interaction Module (CCIM)
sigmoid

Fig. 2: The overall architecture of Row-Column Attention Network.

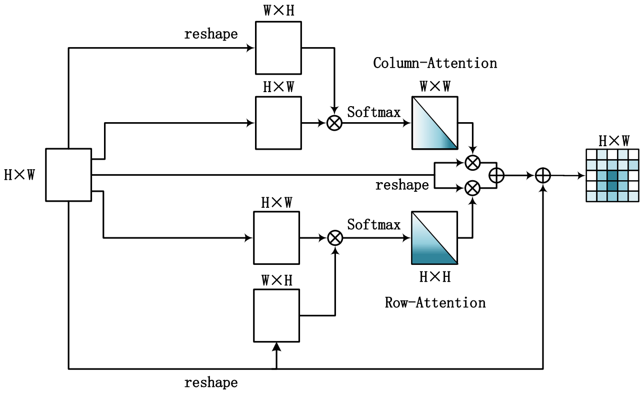

Fig. 3: The details of Row-Column Attention Module.

tion Module (RCAM) and the Cross-Channel Interaction Module (CCIM) for Semantic Segmentation. The visual analysis of our RCANet is shown in Fig. 1, the bar-plot on the right and top shows the correlation between the rows and columns respectively, which are extracted from the row-wise and column-wise attention maps. It can be seen that the row and column have the highest correlation with themselves. And the correlations between the rows and columns containing the same object are much higher than irrelevant rows and columns. Comparing to the attention map generated by the NL module, RCANet is able to capture the global contextual information at different locations more effectively. The contributions of our work can be summarized as:

1. We propose a novel Row-Column Attention Network to implement an efficient self-attention layer which captures the long-distance dependencies not only in spatial dimensions but only across the feature channels.
2. Instead of computing the expensive pixel-level correlations, our method computes the row- and column-level correlations, greatly reducing the computations.
3. We achieve competitive performance against the state-of-the-art methods on Cityscapes and ADE20K.

2. PROPOSED METHOD

2.1. Overall architecture

The RCANet consists of a backbone CNN and a self-attention layer with two modules. The backbone CNN and the self-attention layer will be learned jointly from scratch. The overall architecture is shown in Fig. 2.For an input image, we first compute the feature representation X of shape $C \times H \times W$ by the backbone CNN, where C, H, W are the channel and spatial dimensions respectively. Then we split X into C channels, denoted as $X_1, \dots, X_C$; each channel X_i is of shape $H \times W$. For each channel X_i, we group activations in terms of rows and columns to draw contextual information, which are named as *row attention* and *column attention* respectively. The raw channel X_i can be refined by multiplying itself to the row and column attention maps with a skip connection. The intra-channel operations are referred to as Row-Column Attention Module (RCAM). We then combine the refined channels to form a new feature representation $X^{'}$ of the same shape as X, and construct a Cross-Channel Interaction Module (CCIM) to exchange information between channels, which produces the re-weighting coefficients for each channel. Finally, we re-weight $X^{'}$ and add it to X to yield $X^{''}$.

2.2. Row-Column Attention Module

The Non-local based models [8, 10, 12, 13] use a self-attention mechanism to capture long-range contextual information, the most critical step is to calculate the global attention map $A \in \mathbb{R}^{HW \times HW}$ from the input feature $X \in \mathbb{R}^{C \times H \times W}$. In order to obtain A, one has to compute the correlation between each pair of pixels, each of which involves taking a dot product between two C dimensional vectors, and therefore the time complexity is $\mathcal{O}(CH^2W^2)$. Existing models [8, 10, 12, 7, 13] apply a channel dimension reduction before computing the self-attention, which however does not solve the massive computations over the spatial dimensions. In order to model the global contextual dependencies more efficiently, we introduce the Row-Column Attention Module (RCAM) such that the self-attention on the input X can be done in $\mathcal{O}(CH^2 + CW^2)$ time.

As shown in Fig. 3, the RCAM is applied separately to each channel X_i, which involves the following operations with learnable

Table 1: Comparison with state-of-the-art models on Cityscapes validation set. We only used the fine-data to train RCANet.

Method	Backbone	mIoU(%)
DeeplabV3 [4]	ResNet-101	79.3
ACFNet [14]	ResNet-101	80.08
ANNet [13]	ResNet-101	80.1
CCNet [10]	ResNet-101	81.3
DANet [8]	ResNet-101	81.5
RCANet	ResNet-101	81.8
HRNetV2-W18	HRNetV2-W18	76.2
HRNetV2-W40	HRNetV2-W40	80.2
HRNetV2-W48	HRNetV2-W48	81.1
OCR [15]	HRNetV2-W48	81.6
RCANet	HRNetV2-W48	83.3

coefficients α and β:

$$X_i^{'} = \alpha X_i^{\mathrm{row}} + \beta X_i^{\mathrm{col}} + X_i, \tag{1}$$

where X_i^{row} and X_i^{col} are the outcomes by multiplying the row and column attention maps to X_i respectively. Specifically, we have

$$X_i^{\mathrm{row}} = \underbrace{\mathrm{softmax}\big(X_i X_i^{\top}\big)}_{A^{\mathrm{row}}} X_i \tag{2}$$

$$X_i^{\mathrm{col}} = X_i \underbrace{\mathrm{softmax}\big(X_i^{\top} X_i\big)}_{A^{\mathrm{col}}}, \tag{3}$$

where $\mathrm{softmax}$ is a function to normalize each row of the input matrix to sum to 1.

Intuitively, $A^{\mathrm{row}} \in \mathbb{R}^{H \times H}$ is the row attention map and $A^{\mathrm{col}} \in \mathbb{R}^{W \times W}$ is the column attention map. They encode the correlations between rows or columns. Note that the k-th row in $X_i^{\mathrm{row}}[k, :] = A^{\mathrm{row}}[k, :]X_i$ can be understood as a convex combination of the rows of X_i. Thus, the k-th row of X_i^{row} is formed by fusing highly-correlated rows to the k-th row of X_i. Similarly, the construction of X_i^{col} is done by fusing highly-correlated columns. As such, the output of RCAM $X_i^{'}$ is a refactorization of the input X_i in terms of row and column correlations.

2.3. Cross-Channel Interaction Module

Comparing to the self-attention in Non-local based models, where each element in the attention map is computed using information from all channels, RCAM operates channels independently and computes intra-channel attention maps. To exchange information between channels in a lightweighted fashion, we propose the Cross-Channel Interaction Module (CCIM), which is inspired by SENet [16] and ECANet [17].

Taking as input the output $X^{'}$ from RCAM, the computation in CCIM can be summarized as

$$X^{''} = \mathrm{MLP}\big(\mathrm{GAP}(X^{'})\big) \odot X^{'} + X, \tag{4}$$

where GAP is the global average pooling layer and MLP is a two-layer perceptron with ReLU activation; $\mathrm{MLP}\big(\mathrm{GAP}(X^{'})\big)$ is a C dimensional vector, which will be used to re-weight the channels of $X^{'}$, as denoted by the $\odot$ operation. The architecture of CCIM is illustrated in Fig. 2.

Table 2: Comparison with state-of-the-art models on Cityscapes test set. Note that our model is trained with **only fine annotated data.**

Method	Backbone	mIoU(%)
Model learned on the train set		
PSPNet [18]	ResNet-101	78.4
PSANet [19]	ResNet-101	78.6
PAN [12]	ResNet-101	78.6
AAF [20]	ResNet-101	79.1
HRNetV2-W48 [21]	HRNetV2-W48	80.4
RCANet	HRNetV2-W48	81.56
Model learned on the train+val set		
DeeplabV3[4]	ResNet-101	78.5
PSANet [19]	ResNet-101	80.1
CCNet [10]	ResNet-101	81.4
DANet [8]	ResNet-101	81.5
ACFNet [14]	ResNet-101	81.9
GFFNe [22]	ResNet-101	82.3
HRNetV2-W48 [21]	HRNetV2-W48	81.2
OCR [15]	HRNetV2-W48	82.5
RCANet	HRNetV2-W48	82.64

2.4. Embedding RCAM and CCIM into Networks

To summarize, RCAM concentrates only on refining channels by gathering contextual information within each channel, and CCIM glues individually refined channels back by a re-weighting operation. By chaining these two modules together, we obtain an instance of the self-attention layer, for which we name as the RCA+CCI layer. Comparing to the standard self-attention layer used by the non-local network, RCA+CCI is more suitable for semantic segmentation when the output of the feature extractor (i.e., the backbone) is of high-resolution. On the contrary, the vanilla self-attention layer is infeasible for a backbone architecture with high-resolution outputs due to the $O(CH^2W^2)$ complexity. As an example, a vanilla self-attention cannot be applied to improve HRNet [21] whose output is of shape $720 \times 128 \times 256$ on the Cityscapes dataset [23]. Moreover, RCAM does not require a dimension reduction of the input features, making it more effective in retaining the raw channel information.

3. EXPERIMENTS

3.1. Datasets

Cityscapes has recorded 5,000 finely annotated street scenes images from 50 different cities. ADE20K contains 20,000 images are used for training. There are totally 150 semantic classes in this dataset and the objects in the images are non-uniform distributed, so they are close to natural objects in daily scenes.

3.2. Implementation Details

We use HRNetV2-W48 [21] and ResNet-101 [24] with dilated convolution pre-trained on ImageNet as the backbones of our RCANet. We set the crop size of 512×1024 and 520×520 on Cityscapes and ADE20K respectively. RCANet uses SGD for gradient descent, the Momentum is 0.9 on the whole. The weight decay is 0.0005 for Cityscapes and 0.0001 for ADE20K. The base learning rate is 1e-2 and 2e-2 for Cityscapes and ADE20K, respectively, and learning rate is multiplied by $(1 - \frac{iter}{totaliter})^{0.9}$ after each iteration. We set the weight decay 5e-4 to gradually decrease the learning rate after each iteration. The batchsize is set to 12 for Cityscapes and 16 for

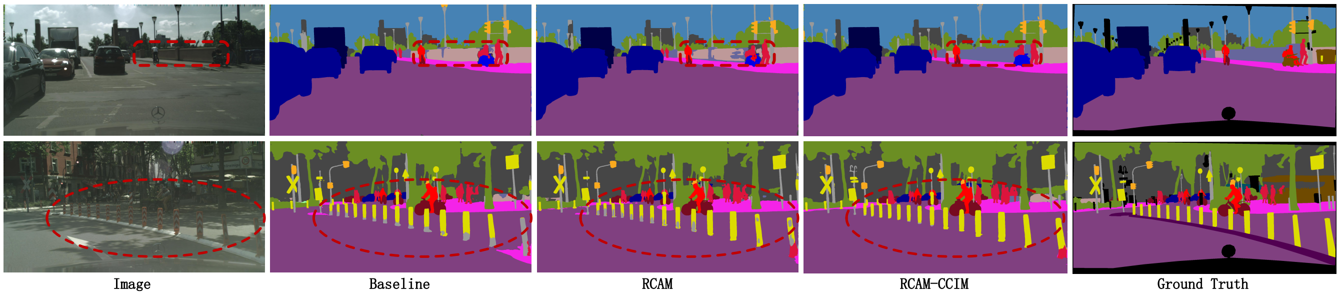

Fig. 4: Visualization the results of baseline, RCAM and RCAM-CCIM on Cityscapes validation set. The red box marks the part where the predicted label changes significantly.

Table 3: Comparsion of GFLOPs and Memory usage estimated under the HRNetV2-W48 backbone with the input size of 1 × 720 × 128 × 256.

Method	GFLOPs	Memory(MB)
Non-local Module [7]	1545.1	10308
Dual-path Module [8]	685.47	6339.88
Criss-Cross Module [10]	236.05	3082.57
OCR [15]	510.1	4036
RCAM-CCIM	38.13	1680

ADE20K. We use multi-scale and random flip for data augmentation in the training process, the epochs of two datasets are 480 and 120 respectively.

3.3. Results on Cityscapes

3.3.1. Compare with state-of-the-arts

We conduct our experiments on Cityscapes validation set and the results are summarized in Table 1. The RCANets use HRNetV2-W48 and ResNet-101 as the backbones for experiments. They both outperform the state-of-the-art 1.7% and 0.3% respectively. When we use HRNet as the backbone, the input of RCAM+CCIM is high-resolution feature maps, which are more conducive to capture global contextual information. Therefore, our RCANet with HRNet can achieve better feature representation to improve the segmentation performance.

The results on the Cityscapes test set are shown in Table 2. We use training set only and train+val sets as the train set, respectively. The process of training only uses fine annotated data, without coarse data and extra datasets. RCANet has reached 81.56% and 82.64% mean IoU on the test set respectively, achieving the state-of-the-art on those two settings. Table 3 shows that our proposed RCANet outperforms existing modules in both computational efficiency and memory cost, which can be applied with only a slight increase in overhead.

3.3.2. Ablation Study

The ablation study is conducted on Cityscapes validation set. To verify the effectiveness of the proposed RCAM and CCIM, we construct different modules and employ them on the top of the baseline. As shown in Table 4, we only use the RCAM can improve the performance by 1.93% against baseline. After we integrate the RCAM and CCIM, the performance can further be improved to 83.3%, significantly increasing the Mean IoU over the baseline by 2.2%.

To further analyze the influence of RCAM and CCIM, we visualize the prediction results produced by baseline, RCAM and RCAM-CCIM respectively. As shown in Fig. 4, RCAM can rectify the error predictions in the baseline through the correlation between the rows and columns.

Table 4: Ablation study on Cityscapes validation set.

Backbone	Method	mIoU(%)
HRNetV2-W48	baseline	81.1
	+RCAM	83.03
	+RCAM-CCIM	83.3

Table 5: Comparison with state-of-the-art models on ADE20K validation set.

Method	Backbone	mIoU(%)
RefineNet [25]	ResNet-101	40.2
PSPNet [18]	ResNet-101	43.29
PSANe [19]	ResNet-101	43.77
EncNet [26]	ResNet-101	44.65
GCUNet [27]	ResNet-101	44.81
CCNet [10]	ResNet-101	45.22
GFFNet [22]	ResNet-101	45.33
RCANet	ResNet-101	45.36
HRNetV2-W48 [21]	HRNetV2-W48	44.2
OCR [15]	HRNetV2-W48	45.5
RCANet	HRNetV2-W48	45.76

3.4. Results on ADE20K

ADE20k is a very challenging semantic segmentation dataset. As shown in Table 5, the performance of RCANet compare to the GFFNet [22] is only slightly improved, because the scene information contained in ADE20K is not conducive to RCAM for extracting sufficient contextual features. Due to the high resolution features provided by HRNet, the RCANet with HRNet as the backbone achieves the best performance, and achieves an improvement of 0.26% compared with HRNet-OCR.

4. CONCLUSION

In this paper, we propose a Row-Column Attention Network (RCANet) for semantic segmentation. We use Row-Column Attention Module (RCAM) to capture global contextual information independently by grouping pixels in rows and columns in each channel. Instead of compute the pixel-level correlations, our method is very efficient. Then we use Cross-Channel Interaction Module (CCIM) to adaptively re-weight the channel information to capture the inter-channel dependencies. Ablation study shows the effectiveness of our RCAM and CCIM. Experiments show our work achieves very promising performance on Cityscapes and ADE20K.

5. REFERENCES

[1] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV 2018*. 2018, vol. 11211 of *Lecture Notes in Computer Science*, pp. 833–851, Springer.

[6] Qihang Yang, Tao Chen, Jiayuan Fan, Ye Lu, Chongyan Zuo, and Qinghua Chi, "Eadnet: Efficient asymmetric dilated network for semantic segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. 2021, IEEE.

[7] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149.

[9] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen, "Axial-deeplab: Standalone axial-attention for panoptic segmentation," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, 2020.

[10] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 603–612.

[11] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1971–1980.

[12] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang, "Pyramid attention network for semantic segmentation," in *BMVC 2018, Newcastle*. 2018, p. 285, BMVA Press.

[13] Zhen Zhu, Mengdu Xu, Song Bai, Tengteng Huang, and Xiang Bai, "Asymmetric non-local neural networks for semantic segmentation," in *ICCV 2019,*. 2019, pp. 593–602, IEEE.

[14] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding, "Acfnet: Attentional class feature network for semantic segmentation," in *ICCV 2019*. 2019, pp. 6797–6806, IEEE.

[15] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Object-contextual representations for semantic segmentation," in *ECCV 2020*. 2020, vol. 12351 of *Lecture Notes in Computer Science*, pp. 173–190, Springer.

[16] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[17] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11531–11539.

[18] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *CVPR 2017*. 2017, pp. 6230–6239, IEEE Computer Society.

[19] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *ECCV 2018*. 2018, vol. 11213 of *Lecture Notes in Computer Science*, pp. 270–286, Springer.

[20] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X. Yu, "Adaptive affinity fields for semantic segmentation," in *ECCV 2018*. 2018, vol. 11205 of *Lecture Notes in Computer Science*, pp. 605–621, Springer.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696.

[22] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, and Kuiyuan Yang, "GFF: gated fully fusion for semantic segmentation," *AAAI,2020*, 2020.

[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR 2017*. 2017, pp. 5168–5177, IEEE Computer Society.

[26] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[27] Yin Li and Abhinav Gupta, "Beyond grids: Learning graph representations for visual recognition," in *NeurIPS 2018*, 2018, pp. 9245–9255.