EasyChair Preprint

№ 2547

Structural and Relational Reasoning with Multi-Scale Context for Semantic Segmentation

Yukihiro Domae, Hiroaki Aizawa and Kunihito Kato

February 4, 2020

Structural and Relational Reasoning with Multi-Scale Context for Semantic Segmentation

Yukihiro Domae[1], Hiroaki Aizawa[1], and Kunihito Kato[1]

[1] Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan
domae@cv.info.gifu-u.ac.jp

Abstract. It is important for semantic segmentation to learn two types of context information. One is global context information for understanding objects, relations between objects, and scenes in input images. The other is multi-scale context information for adapting to changes in the scale and shape of objects. In this research, we tackle the problem of learning to extract them for semantic segmentation. To achieve this, we propose a novel unit that learns to perform structural and relational reasoning by selecting the multi-scale context. The multi-scale context is extracted from receptive fields of different sizes of the backbone network and then is implicitly utilized to improve the global context obtained by GloRe. By the proposed unit, our model allows us to perform structural and relational reasoning for semantic segmentation in complex scenes. We conduct experiments on Cityscapes. In particular, our model achieves the mean IoU score of 73.6, which is 1.1% higher than GloRe. Then, by comparing the prediction between the proposed method and GloRe unit, we confirmed the effect of incorporating two types of context information.

Keywords: Semantic Segmentation, Graph Convolution, Global Reasoning

1 Introduction

Semantic segmentation is a task that assigns a label to each pixel of an image. Recent semantic segmentation approaches have advanced considerably with deep neural networks. However, complex scenes that involve the interaction of multiple objects is challenging. In order to correctly recognize such scenes, it is important to learn to extract global context information and multi-scale context information.

The global context information describes the whole image. It is used to understand the relations between objects, the layout of the objects and the scene in the image. Specifically, it can suppress misunderstanding objects with a similar appearance. Typical methods for extracting it are an approach based on attention mechanism [3] and graph structure [4]. DANet[5] is a method that uses the self-attention mechanism. In this method, the relations are calculated by a dot product of all pairs of elements on a grid and channel. Then, the extraction of features related to each element is performed based on the relations. Therefore, each element can have global information. GloRe is a graph-structured approach for extracting global information. GloRe [1] unit can consider the relations between each feature by learning together with edge

weights in Graph Convolution (GC). Each node after GC has fused features with related nodes. By combining them with input features to GloRe unit, features with global context information are obtained. By adding GloRe unit, the model can effectively extract global context information for various recognition tasks. However, these methods are a lack of multi-scale context information. Therefore, these methods often fail to reconstruct objects with varying scale and detailed boundaries and shapes of objects.

The multi-scale context information deals with the detail appearance and shape of objects. It can be used to recognize small objects, object boundaries, and objects with varying scales. There are two approaches for extracting it: an approach that uses dilated convolution [6] and feature maps extracted from receptive fields of different sizes. The former is a method using dilated convolution in the backbone network (ResNet [7]) that performs feature extraction to enlarge a receptive field without reducing the resolution. Furthermore, by fusing the features of the shallow layer and the features of the deep layer using the skip connection, it becomes possible to extract multi-scale context information. The latter is FCN [8] that utilizes feature maps extracted from receptive fields of different sizes. Detailed information missing during feature extraction is supplemented by using multi-scale features that are outputs from each block of the backbone network. However, there is a lack of global context information, because these methods do not consider the relations between features.

Although recent semantic segmentation approaches have achieved a remarkable performance, it is an open problem to extract and consider two types of context information simultaneously. In this study, we propose a novel unit that extracts them. The unit is composed of two parts: the selection module and the relation module. In the selection module, it selects the context which is useful for relational reasoning from the output with receptive fields of different sizes. The selection allows us to perform relational reasoning while maintaining detail appearance and shape of objects. In the relation module, the features extracted by the selection module are used as a projection matrix for converting a feature map in coordinate space to node features in graph space (or from node features to a feature map) in GloRe unit. Their context information can be extracted by considering the relations between their features by GC. Then, as shown in Fig. 1, the difference between GloRe unit and the relation module lies in the projection matrix between a coordinate space and a graph space.

In the experiment, the proposed unit extracts their context information and achieves higher accuracy compared to GloRe.

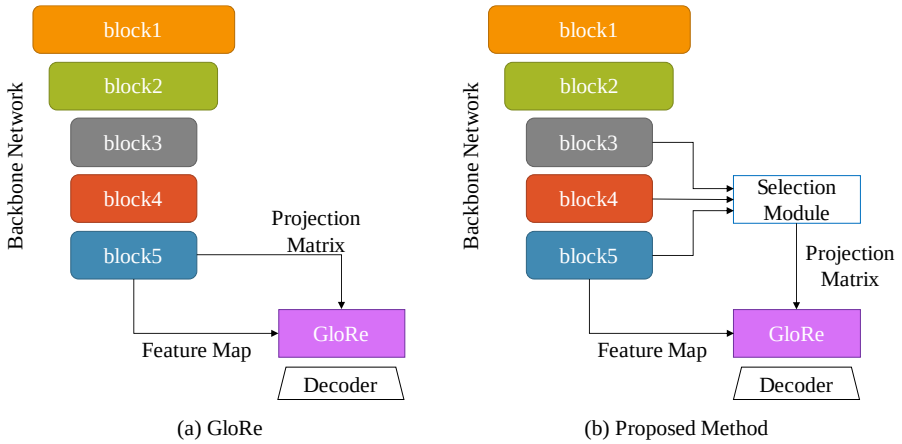

Fig. 1. Comparison between (a) original GloRe and (b) the proposed method. In GloRe, the output of the block 5 in the backbone network is used for the projection matrix from a feature map to node features (or from node features to a feature map) and the input of GloRe. The proposed method uses multi-scale features from the blocks 3, 4, 5 for projection matrix and the output of the block 5 for GloRe input. Then, by learning the relations between selected features obtained from different receptive fields in GloRe, our method can obtain features with two types of context information.

2 Background

We describe the background of the deep learning-based semantic segmentation approach. In early work [8, 9, 10], most models for semantic segmentation employ an encoder-decoder architecture. The architecture consists of an encoder, a decoder, and a classification layer. The encoder extracts the semantic information such as the object and scene from the input image by convolution and pooling layers. The decoder takes the extracted features as input and then reconstructs the spatial resolution of them until that of the input image. The classification layer outputs the probability map based on the upscaled features by a softmax function. As a pioneer work, Fully Convolutional Network (FCN) [8] performs detailed segmentation by combining the prediction result from the deep layer and the shallow layer. SegNet [9] stores the index of each maximum value at each max-pooling layer in an encoder. The decoder maps each element of the feature map to the corresponding position by reusing it, which is called unpooling. As a result, the convolution following unpooling reconstructs coarse features into dense features while considering of the mapped positions and features. However, since the features after unpooling are coarse, there is a limit in performing more detailed segmentation.

To solve the above problems, U-Net [11] predicts segmentation maps by restoring the resolution while combining the features of the encoder whose resolution matches at each level of the decoder. The information before downsampling was supplemented by a path such as a skip connection in encoder and decoder. Therefore, U-Net allows

us to perform more detailed prediction. As described in section 1, the dilated convolution [6] model uses its convolution in a deep layer in the backbone network (ResNet [7]). As a result, there is no need to perform excessive pooling and a decrease in resolution can be suppressed. Furthermore, the features in the shallow layer and the deep layer are combined by the skip connection in ResNet.

As a recent segmentation model [12,13,14,15,16], the FCN-based method in which ResNet with dilated convolution is used as the backbone network has been proposed. FCN-based approaches commonly extract context information using feature maps in the middle or at the end of the backbone network. As an example, a module or unit for extracting global context information such as DANet [5] or GloRe [1] as described in Section 1 is often added. It is possible for ResNet with dilated convolution to extract multi-scale context information due to the following two points. One is dilated convolution makes it possible to expand receptive fields without reducing resolution. The other is multi-scale information can be extracted by combining features of a shallow layer and a deep layer using skip connection. However, there is a problem that detailed boundary information of an object or information of a small object is lost due to a dilation rate of a dilated convolution. In this work, we introduce a novel unit based on GloRe unit to extract two types of context information.

GloRe Unit. The purpose of GloRe unit is to extract global context information for semantic segmentation and classification models. Fig. 2 shows the overview of GloRe. It takes as input the feature map obtained from the backbone network and then outputs a feature map with global context information. First, two pointwise convolutions ($\theta(\cdot)$ and $\phi(\cdot)$) are performed on the input feature map $\boldsymbol{X} \in \mathbb{R}^{C \times L}$ to GloRe unit. C is the number of channels in the feature map. H and W indicate the height and width, respectively. L is the number of spatial elements ($L = H \times W$). The purpose $\theta(\cdot)$ is to obtain the features to consider relations. It of $\phi(\cdot)$ is to reduce the dimensions of a feature map. The outputs from their convolutions are $\theta(\boldsymbol{X}) \in \mathbb{R}^{N \times L}$ and $\phi(\boldsymbol{X}) \in \mathbb{R}^{C' \times L}$ respectively. N is the number of features to be considered for the relations and is also the number of nodes in a graph. Therefore, $\theta(\boldsymbol{X})$ can be considered as a projection matrix for converting the feature map in the coordinate space into node features in the graph space. As shown in equation (1), the feature in the coordinate space is converted into the node features $\boldsymbol{V} \in \mathbb{R}^{C' \times N}$ in graph space. It is considered as applying weighted global average pooling to $\phi(\boldsymbol{X})$ using each of the N features in $\theta(\boldsymbol{X})$ as a weight.

$$\boldsymbol{V} = \frac{1}{N} \phi(\boldsymbol{X}) \theta(\boldsymbol{X})^T. \tag{1}$$

A fully-connected graph is constructed from the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ and the node features $\boldsymbol{V}$. The relations between the N features of $\theta(\boldsymbol{X})$ are learned by Graph Convolution (GC) with $\boldsymbol{V}$ input. GC is as shown in equation (2).

$$\boldsymbol{V}' = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{V}\boldsymbol{W}, \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ is the parameter of GC, and $\boldsymbol{V}' \in \mathbb{R}^{C' \times N}$ indicates the node features after GC. The identity matrix $\boldsymbol{I}$ is a residual pass that alleviates the difficulty of optimization. We randomly initialize $\boldsymbol{A}$ and $\boldsymbol{W}$. Therefore, it is possible to learn the strength of the relations between each node feature by learning the adjacency matrix $\boldsymbol{A}$. The input $\boldsymbol{V}$ of the GC and the output $\boldsymbol{V}'$ of the GC have the same number of nodes and each node of the two graphs has a correspondence. Therefore, as in Equation 3, the weighted broadcast of each corresponding node features $\boldsymbol{V}'$ is performed using each of the N channels of $\theta(\boldsymbol{X})$ as weights and the graph is returned to the coordinate space. Moreover, it expands to the same dimension as the input of GloRe unit by pointwise convolution.

$$\boldsymbol{X}' = f\left(\boldsymbol{V}'\theta(\boldsymbol{X})\right), \tag{3}$$

where $f(\cdot)$ is pointwise convolution. Each node after the GC has fused feature with the related node features. When the fused node features are returned to the feature in the coordinate space, a feature map with global context information is produced by combining the feature before inputting to GloRe unit with the fused feature.

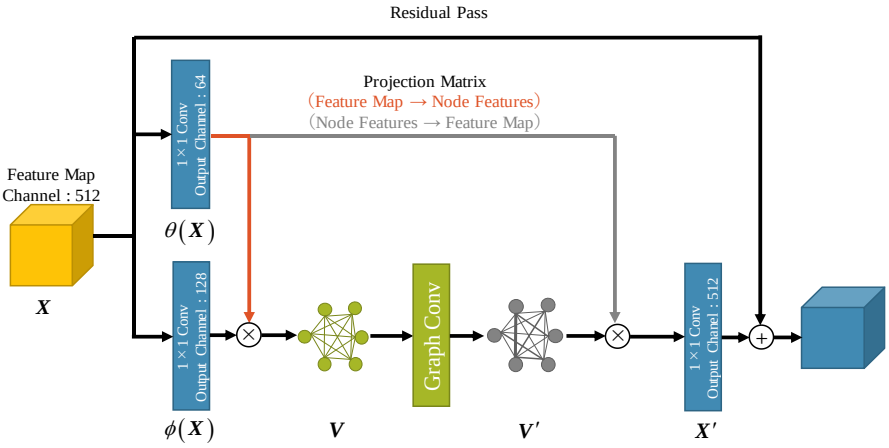

Fig. 2. Overview of GloRe unit. $\boldsymbol{X}$ is input features to GloRe unit. $\theta(\boldsymbol{X})$ is a projection matrix from a feature map to node features (from node features to a feature map). $\phi(\boldsymbol{X})$ is a feature map with a reduced dimension. $\boldsymbol{V}$ is node features that input to GC. $\boldsymbol{V}'$ is node features that output from GC. $\boldsymbol{X}'$ is a feature map after projection from node features to a feature map.

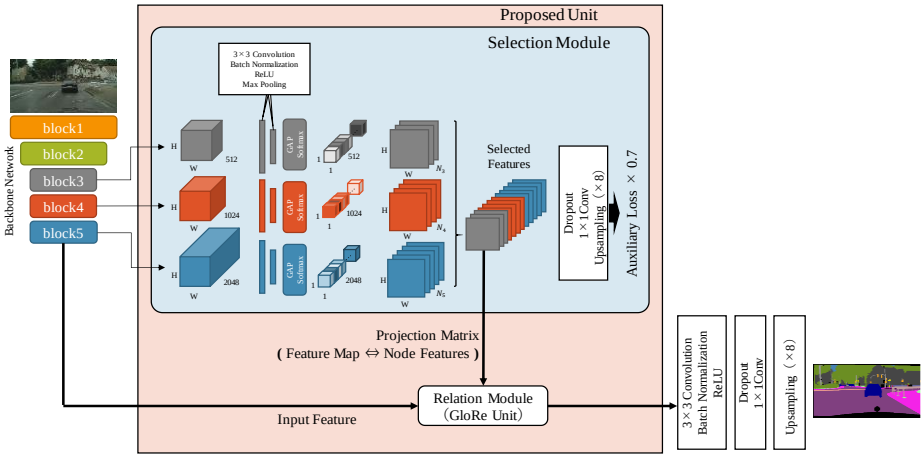

Fig. 3. Overview of our proposed unit. The selection module extracts multi-scale context information and the relation module extracts global context information. The proposed unit outputs a feature map with two types of context information. *H* and *W* are the height and width of the output from each block, respectively. N_3, N_4, and N_5 are the number of features to select from the output of each block.

3 Proposed Method

The purpose of our proposed unit is to extract a feature map with both multi-scale context information and global context information. As shown in Fig. 3, our proposed unit is composed of two modules, a selection module and a relation module. The selection module selects features that accurately capture objects in the image such that they are effective for semantic segmentation from the outputs of each block of the backbone network. In the relation module, GC fuses related information from each multi-scale features selected in the selection module. Hence, a feature map including two types of context information is produced.

3.1 Selection Module

The selection module selects features that are effective for semantic segmentation from the output in each block of the backbone network. The block 3 of the selection module in Fig. 3 is explained as an example. First, convolution and pooling are performed on the features of the block 3 and the features are compressed. Then it applies global average pooling and softmax. The dimension of output from softmax is 1 × 1 × k, where k is the number of channels in a feature map from the block 3. Next, only N_3 features of the block 3 are selected in descending order of softmax value. By performing this operation on each feature map that output from the blocks 3, 4, and 5 of the backbone network, multi-scale context information is obtained. (The number of features selected by the block 4, 5 is N_4, N_5)

3.2 Relation Module

The relation module uses GloRe unit to obtain a feature map with two types of context information using the multi-scale features selected by the selection module. In this module, the multi-scale features are used as the projection matrix between a coordinate space and a graph space. Therefore, the multi-scale features selected by the selection module are converted to node features and the relations between their features are learned by GC. Each node after GC has features obtained by fusing some node features related to the node. Then, when converting from node features to a feature map, each node features are weighted and broadcast using the corresponding multi-scale features as weight. Therefore, a feature map with two types of context information is produced. A detailed prediction of the boundary, shape, and layout is possible while considering global information because the class label of each pixel is predicted based on the features with the two types of context information.

3.3 Auxiliary Loss

Multi-scale features selected from each block in the selection module are combined to predict for semantic segmentation maps. The value of softmax is learned to be so large that it is a feature that accurately captures the object. Therefore, it is possible to extract multi-scale context information that accurately captures an object that is effective for semantic segmentation by the selection module. The optimized loss function L_{opt} is as shown in equation (1).

$$L_{opt} = L + \alpha L_{aux}, \tag{4}$$

where the final prediction loss L and auxiliary loss L_{aux}. Both of them use the cross-entropy loss for semantic segmentation. The term α is set a penalty in L_{aux} to emphasize the loss of the final prediction

4 Experiment

4.1 Experimental Settings

We validate that the proposed method can extract two types of context information. Same with GloRe [1], the backbone network is ResNet-50 [7] with dilated convolution pre-trained by ImageNet [17]. Therefore, the output stride is 1/8. The number of features selected from the blocks 3, 4, and 5 in the backbone network is 11, 21, and 32, respectively. The proposed method is compared with the original GloRe [1] and FCN [8] (without GloRe). We use stochastic gradient descent with momentum 0.9 and weight decay $1e-4$. The scheduling used "ploy" of power 0.9, which is the same as [1]. The initial learning rate was set to 0.006, and the batch size is set to 26. We evaluate the proposed method on Cityscapes [2] (fine annotations). The dataset includes 2975, 500, and 1525 images for the training, validation, and test sets, respectively. We resized the original size of 1024 × 2048 to the fixed size of 512 × 1024. In

the training, we cropped it at the center. The output size of the crop is 512 × 512. In the evaluation, we upscale the predictions to the size of 1024 x 2048 for calculating the IoU score on the test server. The auxiliary loss, which is a feature extracted from the selection module, was suppressed by multiplying it by 0.7 to emphasize the loss in the final prediction.

4.2 Results

Table 1 shows the results of IoU of each class and Mean IoU for each model. Our proposed method was superior to the conventional method in IoU of most classes and mean IoU. Fig. 4 and 5 show the predictions of GloRe and the proposed method. In the yellow boxes of Fig. 4, the small riders and narrow poles have disappeared in GloRe. On the other hand, we confirmed that the proposed method can perform recognition without loss of that features by extracting multi-scale context information from the shallow layer of the backbone network. The objects (wall vs. fence, truck vs. car, and bus vs. train) in yellow boxes of Fig. 5 have similar appearances. In GloRe, ambiguous recognition occurred in the object region. We confirmed that the proposed method can suppress such misclassification by using features that more accurately capture the objects selected by the selection module as the projection matrix in the relation module. Therefore, we confirmed the effect of collecting two types of context information.

Table. 1. Quantitative results on the Cityscapes test set.

	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	Mean IoU
FCN (Without GloRe)	97.6	80.0	89.6	44.6	45.6	45.6	60.8	66.4	91.1	68.3	93.5	77.2	57.2	92.9	51.7	59.4	50.4	58.3	67.1	68.2
Original GloRe	98.2	82.8	91.0	48.7	51.1	51.1	65.6	70.2	91.8	**70.1**	94.1	80.0	64.0	94.1	**59.8**	71.8	62.1	62.0	69.7	72.5
Ours	**98.3**	**83.5**	**91.3**	**50.3**	**51.5**	**52.1**	**66.5**	**71.0**	**91.9**	69.1	**94.2**	**80.3**	**65.0**	**94.6**	58.8	**77.6**	**70.0**	**62.8**	**70.4**	**73.6**

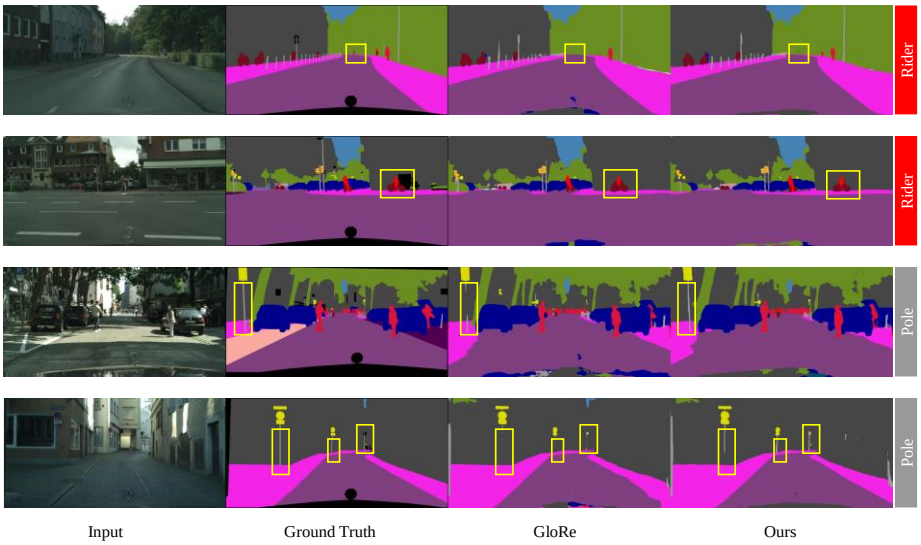

Fig. 4. The comparison with GloRe and our proposed method. Each yellow box shows the effect of extracting multi-scale context information in the proposed method.

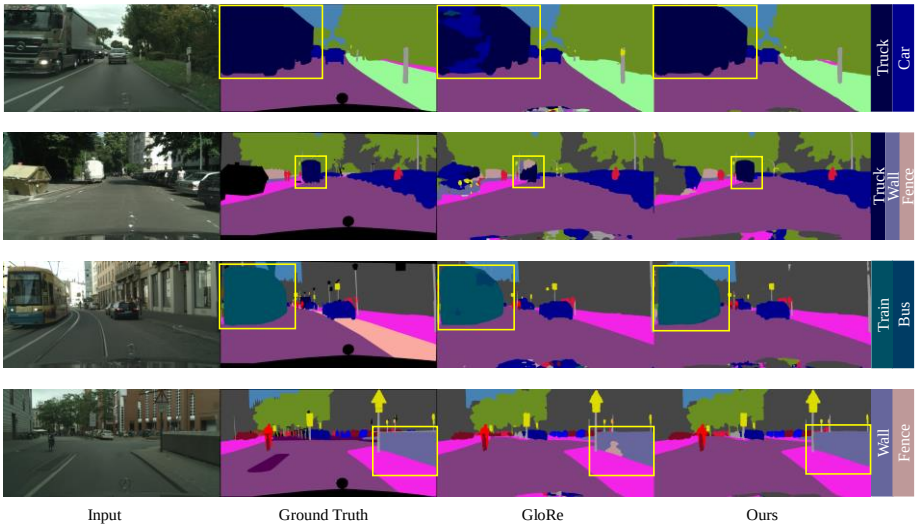

Fig. 5. The comparison with GloRe and our proposed method. Each yellow box shows the effect of extracting both multi-scale and global context information in the proposed method.

5 Conclusion

It is important for semantic segmentation, which performs pixel-wise prediction, to utilize two types of information: multi-scale context information and global context

information. The former is necessary to deal with various scales of the objects. The latter is necessary to understand the relations between objects, the layout of the objects, and the scene in the image. In this research, we proposed a novel GloRe-based unit that learns the relationships between features with multi-scale context. It is useful for refinement of the global context for semantic segmentation. As a result, it became possible to suppress ambiguous recognition in the original GloRe. One of the limitations of this study is that tuning of the penalty term in Auxiliary Loss and the number of selected features in the selection module is sensitive to the performance. Therefore, our model is slightly difficult to apply to other tasks or existing methods. As future work, we aim to resolve the issues and extend it to a versatile unit that can be easily incorporated into existing methods.

References

1. Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y.: Graph-based global reasoning networks. In: The IEEE Conference on Computer Vision and Pattern Recognition, 433-442 (2019).
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R. Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: The IEEE conference on computer vision and pattern recognition, 3213-3223 (2016).
3. Vaswani, A, Shazeer, N, Parmar, N, Uszkoreit, J, Jones, L, Gomez, A.N., Kaiser, L, Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems, 5998–6008 (2017)
4. Kipf, T. N., & Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings, 1–14 (2019).
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition, 3146-3154 (2019).
6. Yu, F., & Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings (2016).
7. He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. In: The IEEE conference on computer vision and pattern recognition, 770-778 (2016).
8. Long, J., Shelhamer, E., & Darrell, T.: Fully convolutional networks for semantic segmenta-tion. In: The IEEE Conference on Computer Vision and Pattern Recognition, 3431-3440 (2015).
9. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In: IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495 (2017).
10. Guosheng, L., Anton, M., Chunhua, S., Ian, R.: "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." In: The IEEE Conference on Computer Vision and Pattern Recognition. 1925–1934 (2017)
11. Ronneberger, O., Fischer, P., & Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, 234-241 (2015)

12. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J.: Pyramid scene parsing network. In: The IEEE Conference on Computer Vision and Pattern Recognition, 2881-2890 (2017).
13. Zhao, H., Zhang, Y., Liu, S., Shi, J., Change Loy, C., Lin, D., & Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: The European Conference on Computer Vision (ECCV) 267-283 (2018).
14. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W.: Ccnet: Criss-cross attention for semantic segmentation.: In: The IEEE International Conference on Computer Vision, 603-612 (2019).
15. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., & Agrawal, A.: Context encoding for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition, 7151-7160 (2018).
16. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., & Lu, H.: Adaptive context network for scene parsing. In: The IEEE International Conference on Computer Vision, 6748-6757 (2019).
17. Krizhevsky, A., Sutskever, I., & Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, 1097-1105 (2012).