2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM) | 978-1-6654-9858-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/MLCCIM55934.2022.00083

A Survey of Semantic Segmentation Methods in Traffic Scenarios

1st Yuanhui Guo
State Key Laboratory of Automotive Simulation and Control
Jilin University
Changchun, China
1592077424@qq.com

2nd Biao Yang*
School of Microelectronics and Control Engineering
Changzhou University
Changzhou, China
yb6864171@cczu.edu.cn

***Abstract*—Semantic segmentation has always been a very challenging research topic in computer vision and deep learning and has extensive applications in real-life scenarios. With the development of computing hardware and deep learning technology, researchers have a higher research enthusiasm for semantic segmentation. This work briefly introduces several semantic segmentation models, datasets, and the main issues of traditional semantic segmentation. Afterward, the current mainstream semantic segmentation algorithm and the experimental results were compared. Finally, future research directions of semantic segmentation are presented and discussed.**

Keywords-component; semantic segmentation; deep learning; computer vision; vision in transformer

I. INTRODUCTION

With the development of equipment performance in computers, deep learning technology is widely used in real-life applications. Since Alex's Alexnet network model[16] has won the championship in the ImageNet2012 image classification competition, the effect of deep learning technology in computer vision is also more apparent. At the same time, CNN(convolutional neural networks) take the advantage in the fields of image segmentation, object detection and super-resolution. In recent years, deep learning technology has dominated most computer vision fields, and semantic segmentation technology is the most critical technology in computer vision. Thanks to the layered screen information provided by semantic segmentation, computers can make better judgments on real vision based on this information.

In real-life scenarios, semantic segmentation results have a good application in agriculture, transportation, medicine and other fields, as shown in Fig. 1. However, semantic segmentation also encountered some challenges in practical applications. The following issues are summarized in the text for the semantic segmentation scenes in traffic scenarios. (1) Dataset problem: For semantic segmentation tasks, the required annotation data is pixel-based, which contains the contour and classification of each object. Therefore, it is difficult to observe the boundaries of each target and distinguish them when manually labeling. This means that people need to pay a lot of time cost for each label. (2) Segmentation accuracy problem: The current network model still has accuracy problems in the effect of semantic segmentation. The pixel-level semantic segmentation task not only needs to distinguish the types of objects, but also needs to distinguish the boundaries of the objects. Therefore, semantic segmentation tasks are more difficult than traditional image classification and object detection tasks. There are also many researchers trying to improve the accuracy of semantic segmentation. (3) Computer performance problems: in the semantic segmentation scene, in order to achieve pixel-level classification effects, we often need to design a deeper network model. The application in traffic scenarios usually considers real-time problems. Therefore, how to design a small resource overhead and better effect at the same time are also essential issues in semantic segmentation. This work describes the reasons for the above problems and then summarizes the current mainstream solutions, and looks forward to subsequent research and development.

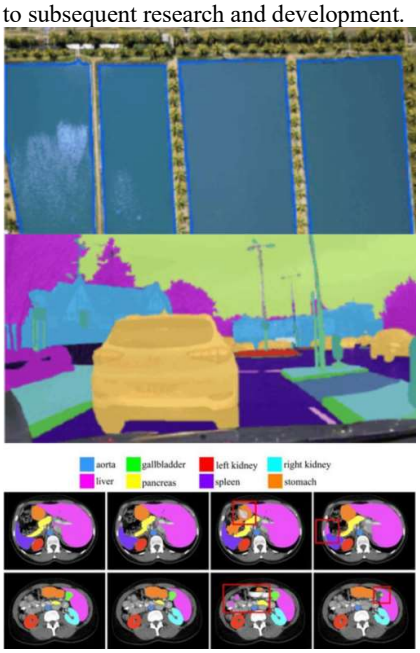

Figure 1. Application of semantic segmentation. From top to bottom are applications in agricultural, transportation, and medicine.

II. BACKGROUND

A. Traditional semantic segmentation algorithm

In traditional image segmentation algorithms, image segmentation is often regarded as a mathematical solution. In the past, some scholars[14] used the optimization of the objective function to solve the image segmentation problem. However, the calculation of such schemes suffers from heavy computing burdens. Besides, it can only be used for images of similar classes. Thus its generalization is not strong. Other scholars[15] divide images through edge detection algorithms. The images are smooth through the filtering algorithm, and then the edge details of the image edges are detected through convolution and gray perception. Although this type of algorithm overcomes the shortcomings of extensive calculations in the past, it cannot

978-1-6654-9858-6/22/$31.00 ©2022 IEEE
DOI 10.1109/MLCCIM55934.2022.00083

ensure the continuity and closure of the edges, and it performs poorly in the details. In recent years, Deep learning techniques and computer performance are developing rapidly. Scholars have applied deep learning technology to image segmentation. These methods include CNN, activation functions, pooling operations, etc.. It is convenient to extract feature information from the image to achieve semantic segmentation. In 2015, Long et al.[5] proposed FCN(fully convolutional network), Fully Convolutional Network (FCN), which uses pooling and fully connected layers to process the final features and predict the class of each pixel. The network framework is shown in Fig. 2. The FCN expands the image classification task to the image segmentation field so that the image segmentation is combined with the classification algorithm. Since then, deep learning algorithms take advantage in semantic segmentation. More and more scholars are engaged in research in this field.

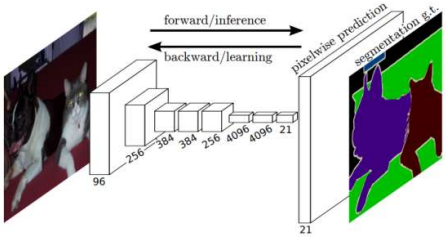

Figure 2. The architecture of a fully convolutional network[7].

B. Semantic Segmentation Dataset

Semantic segmentation datasets also have different types. Commonly semantic segmentation datasets include panoramic images, RGB-D depth images, 3D point cloud images, and RGB two-dimensional images. Generally, the semantic segmentation images in traffic scenarios are two-dimensional, such as Cityscapes, GTAV and CamVid[6], as shown in Fig. 3.

Figure 3. Commonly used Semantic Segmentation Dataset. From left to right are original RGB maps and semantic segmentation results, respectively. From top to bottom are Cityscapes, GTAV, and CamVid[6], respectively.

1) Cityscapes: The scene of the Cityscapes dataset is taken from urban traffic scenarios, which contains 5,000 meticulously labeled and 20,000 crudely labeled. These labels comes from 50 different cities, including different seasons of spring, summer, autumn, and good background changes. It is a rich dataset in traffic scenarios.

2) GTAV: The GTAV dataset is different from other datasets. The images is not come from the real world but from a game which called GTAV(Grand Theft Auto V). The author of the dataset achieves a very accurate pixel-level semantic segmentation label through interface calls in the game. The city view of the dataset is the same as the American-style city. The image of the dataset is taken from the traffic scenario in the game, including 24,966 finely labeled images. The dataset also include 19 categories such as roads, buildings, sky, sidewalks, cars, etc.

3) CamVid: CamVid dataset came to the driving traffic scenario of the British city. Its images came from the actual shooting of the driving vehicle in the traffic scenario, so the images were closer to the actual distribution of traffic scenarios. The dataset contains 32 categories, including roads, vehicles, buildings, and pedestrians. It is divided into 367 training images, 100 validation images and 233 test images. More importantly, all its data are from traffic scenarios. This means that CamVid has an important role in real-time semantic segmentation and intelligent driving.

C. Semantic segmentation evaluation standards

Semantic segmentation algorithm is the same as other computer algorithms. It has specific evaluation standards. The commonly used semantic segmentation evaluation standards include inference, parameters, and MIOU. Here are three commonly used evaluation indicators.

1) Inference time: Inference time refers to the segmentation result of the target image from an image in the calculating device, which is related to the computer's and GPU's performance. Therefore, when the semantic segmentation inference evaluation is performed, the same device is generally used to compare the inference time of a particular algorithm, and the unit is a millisecond.

2) Parameter scale: In the algorithm of deep learning, the parameter is also a critical evaluation standard. Due to the different memory sizes of the GPU, smaller parameters can better adapt to lightweight edge equipment. If the equipment is in the driving scenario, its GPU's memory size is generally small, so a tiny scale of parameters can better adapt to more equipment.

3) MIOU: In semantic segmentation, prediction accuracy is a relatively important evaluation standard, and the error caused by the accuracy will bring uncertainty in practical applications. MIOU is generally used as an evaluation indicator in semantic segmentation. This is the calculation of MIOU.

$$MIoU = \frac{1}{c+1}\sum_{i=0}^{c}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \quad (1)$$

where c represents the number of categories, p represents pixels, i represents the real value of prediction, j represents the predicted value. p_{ij} denotes that i is predicted as j and vice-versa.

III. PRECISE SEMANTIC SEGMENTATION

In applying semantic segmentation, accuracy is an essential part of semantic segmentation. More accurate semantic segmentation models can bring better segmentation effects, reducing unnecessary processing due to poor image segmentation effects. In recent years, scholars have obtained higher accuracy through different schemes to achieve higher accuracy, mainly including ViT (Vision in Transformer) and parallel multi-channel network architectures. It adopts a multi-branch design on the network, and different branches perform different segmentation tasks to achieve higher segmentation effects. ViT resorts to the Transformer architecture to conduct network inference through operations such as segmentation, encoding, and decoding. Thanks to TansFormer's efficient feature extraction ability, ViT has achieved good results in semantic segmentation. The following two types of networks are introduced as follows.

A. Multi-Channel Network Architecture

In 2018, Yu et al.[7] proposed the multi-channel segmentation network in semantic segmentation. They proposed a new type of Bilateral Segmentation Network (BiSeNet). One is a spatial perception branch, which retains more spatial information through smaller steps, to better perceive the spatial resolution features. The other obtains richer context information through a higher multiple of down-sampling and attention concentration mechanism. Finally, semantic segmentation is achieved by integrating two features. The architecture of BiSeNet is shown in Fig. 4.

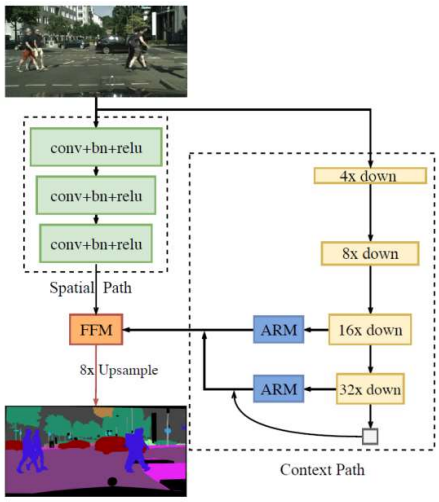

Figure 4. Bilateral Segmentation Network Architecture[7].

As shown in Fig. 4, Bilateral Segmentation Network Architecture has two branches responsible for different tasks. The spatial perception branch obtains more spatial information through the smaller step-long convolution. The context branch uses larger down-sampling ratios and the attention refinement module (ARM). ARM quickly extracts the context information using global pooling, normalization, and constant connections. The larger down-sampling ratios also improve the ability of context information extraction. Finally, features are combined through a feature fusion module (FFM). BiSeNet has achieved SOTA effects on datasets such as Cityscapes and CamVid.

Since then, the multi-channel network has become the basic architecture for obtaining higher MIOU in the semantic segmentation methods that use CNN. In 2020, Yuan et al.[8] applied the atrous convolution to the spatial network and achieved SOTA effects on multiple datasets. In the same year, they made SOTA effects through contextual relationships between polymerization pixels. These networks achieve great results in lightweight real-time semantic segmentation

B. Vision in Transformer (ViT)

In 2017, Vaswani et al.[9] proposed the Transformer to process natural language text. It improves the convergence issue in training a recurrent neural network (RNN), and achieves rapid parallel running through self-attention. Although researchers have tried to apply Transformer into vision, its semantic segmentation performance is unsatisfactory. Until 2021, Liu et al.[10] proposed a hierarchical Swin-Transformer by changing the operations of fixed dividing the image window. The sliding windows contain non-overlapping local windows and overlapped cross windows. The calculation of attention is limited to one window, introducing the locality of convolution operation and saving computation. The difference in the sliding window between Swin-Transformer and ViT is illustrated in Fig. 5.

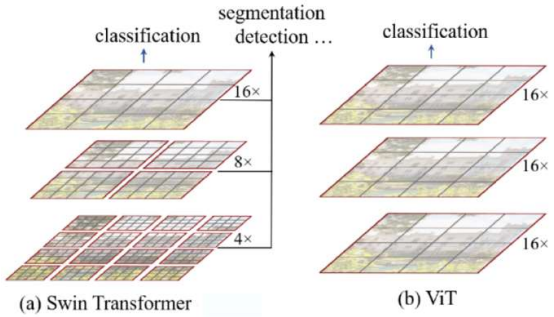

Figure 5. The difference of sliding window between Swin-Transformer and ViT[10].

Swin-Transformer is different from traditional ViT in design. Its encoding in embedding is optional, whereas its encoding in self-attention is not optional and is a relative position encoding. Traditional ViT generally adds a learning parameter as a classification token, while Swin-Transformer directly averages and outputs the results of classification. The architecture of the Swin-Transformer is shown in Fig. 6. Swin-Transformer performs excellently in semantic segmentation. It surpasses the previous semantic segmentation methods in cityscapes, ADK, and CamVid.

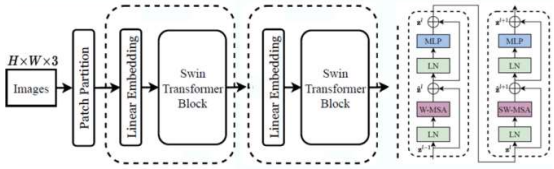

Figure 6. Illustration of the Swin-transformer architecture[10].

IV. Real-Time Semantic Segmentation

Although semantic segmentation in agriculture and medicine does not need real-time performance, traffic scenarios such as autonomous driving are likely to cause traffic accidents without real-time calculations. Recently, researchers generally designed some detailed architectures of different CNN to achieve faster reasoning speed and better feature extraction. The excellent designs include atrous convolution and self-attention.

A. Atrous Convolution

Before 2015, the atrous convolution was rarely used in classification tasks because it is poor in image details. With the rise of pixel-level semantic segmentation, atrous convolution rapidly develops because it makes the model learn the spatial characteristics of the target faster and reduces the computing burdens. The illustration of atrous convolution is shown in Fig. 7.

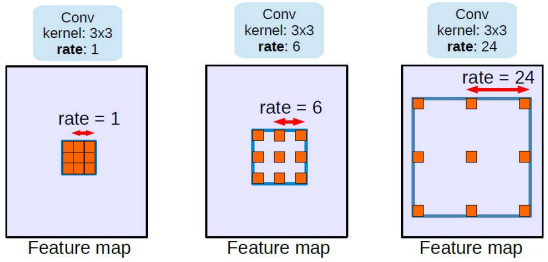

Figure 7. The illustration of atrous convolution with different rates[11].

The applications of atrous convolution in semantic segmentation are mainly from DeepLab[11]. DeepLabv1 uses atrous convolution to obtain more dense and robust feature maps. However, the resolutions of obtained feature maps are too low. DeepLabv2 avoided the resolution loss of the feature diagram by modifying the final pooling operations. Afterward, DeepLabv3 re-discussed the use of atrous convolution. Under the framework of serial modules and spatial pyramids, it can get more remarkable experience to get multi-scale features. DeepLabv3 can get a good MIOU on various datasets. Therefore, the operation of the atrous convolution has become the main functional module in the semantic segmentation models that use CNN.

B. Self-Attention

In 2015, natural language processing (NLP) introduced deep learning insights. However, the features of language text are very abstract, and deep learning algorithm is difficult to extract the characteristics of natural language. Most deep learning algorithms are challenging to converge in NLP. To solve this problem, some scholars have proposed self-attention. Since then, self-attention has been introduced to object detection, semantic segmentation, and other fields. In semantic segmentation, commonly used self-attention mechanisms consist of spatial attention, channel attention, and spatial-channel self-attentions.

Squeeze-and-Excitation Networks (SeNet)[12] won the 2017 ImageNet championship. The authors superimposed the weight of each channel based on the original deep CNN, making it easier for the network to pay attention to different channels. SeNet contains the following steps: (1) The features extracted by the network are input to two fully connected layers, and then input to the global average pooling (2) A Sigmod activation is used to calculate the attention weight from the network features, and finally output is generated by multiplying the network features with the attention weight. CBAM[13] argues that only using the channel attention module will lose the context features. Therefore, it combined channel attention with spatial attention. Both attention mechanisms are shown in Fig. 8.

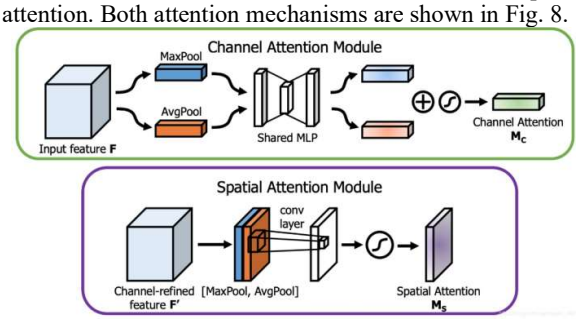

Figure 8. Spatial and channel attention module[13]

Since then, self-attention has a great advantage in computer vision, such as DeepLab, PIDNet, EfficientNet. The self-attentions are used to make the network more convergence during training and speed up the inference process.

V. Performance Test

The comparisons of different semantic segmentation methods should comprehensively consider their inference duration, parameter scales, and training data. In this work, The Cityscapes dataset is used to evaluate several network models proposed in recent years. The comparison results, including whether using pre-training, FPS, parameter scales, and MIOU, are reported in Table 1.

It can be concluded from Table 1 that recent research can be divided into three types: (1) Only the evaluation indicators of MIOU are considered, which aim to improve the performance of semantic segmentation, such as Vit-Adapter-L, SWIN-L, and Lawin-L. (2) Considering more indicators for reasoning, such as PP-Liteseg, and Deeplabv3. (3) Comprehensive consideration of the two, such as PIDnet, BiSeNet, and HRNETV2.Then, this paper conducts a semantic segmentation test on PP-liteseg, STDC2-Seg and PIDNet. The test results are shown in Figure 9.

VI. Summarize

This work first introduces the semantic segmentation datasets from the semantic segmentation in traffic scenarios. Afterward, several accurate semantic segments based on multiple branches and VIT architectures are discussed. Subsequently, several semantic segmentation networks focusing on real-time concentrations and self-attentions are introduced. Finally, evaluations of discussed methods on the Cityscapes dataset are performed. According to the evaluation results, the future semantic segmentation network should focus on two aspects: (1) The accuracy of semantic segmentation, especially for applications in medicine and others without real-time requirements. (2) The combination of accuracy and reasoning speed, especially for applications in intelligent transportations that call for accuracy and speed. (3) The small-scale semantic segmentation datasets need unique solutions to reduce the data dependence.

ACKNOWLEDGMENT

This work is supported by Foundation of State Key Laboratory of Automotive Simulation and Control with NO. 20210241; Postdoctoral Foundation of Jiangsu Province NO. 2021K187B; National Postdoctoral General Fund NO. 2021M701042; General project of Jiangsu Scicence and Technology Department NO.BK20221380.

TABLE I. COMPARISON RESULTS OF DIFFERENT SEMANTIC SEGMENTATION RESULTS ON CITYSCAPES (RESULTS ARE CITED FROM THEIR ORIGINAL WORKS)

Net	pre-train	FPS	Param	MIOU(%)	year
FCN(VGG16)	Yes	10.2(NVIDIA Titan X)	134.0M	65.3	2016
BiSeNet(Xception39)	No	105.8(NVIDIA Titan XP)	20.9M	68.4	2018
HRNetV2 + OCR	Yes	22.2(NVIDIA P40)	105.0M	84.5	2020
Deeplabv3+(mobilenet)	Yes	72.5(NVIDIA P40)	5.8M	76.9	2018
ViT-Adapter-L	Yes	18.5(NVIDIA V100)	571.0M	85.2	2022
Swin-L	Yes	42.1(NVIDIA V100)	197.1M	83.2	2021
Lawin-L	No	11.8(NVIDIA V100)	201.2M	84.4	2022
PIDNet-M	Yes	31.1(NVIDIA 3090)	120.0M	79.8	2022
PIDNet-L	Yes	42.2(NVIDIA 3090)	173.0M	80.6	2022
PP-LiteSeg	Yes	273.0(NVIDIA 1080ti)	14.5M	72.0	2022
BiSeNetv2(STDC1-Seg50)	Yes	250.4(NVIDIA 1080ti)	8.4M	71.9	2021
BiSeNetv2(STDC2-Seg50)	Yes	188.6(NVIDIA 1080ti)	12.5M	73.4	2021

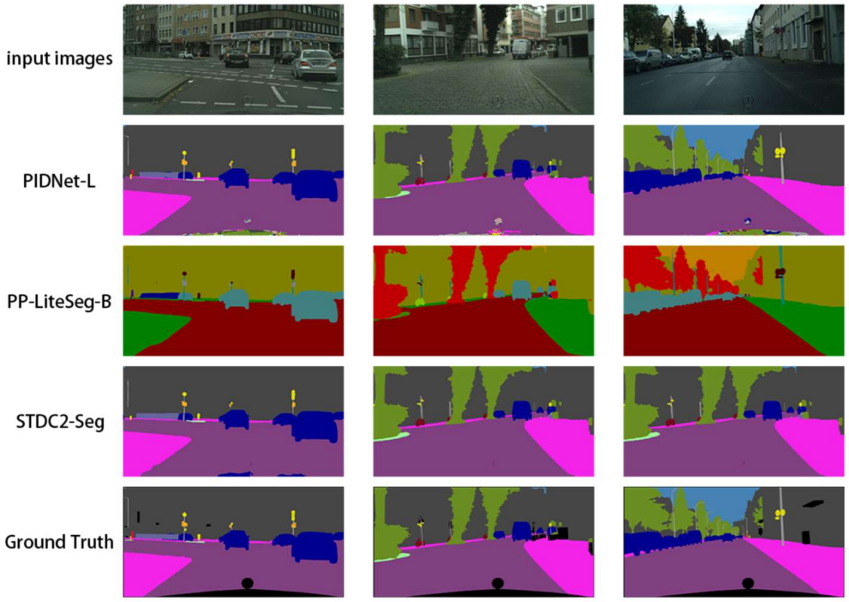

Figure 9. Comparison result of different semantic segmentation results on cityscapes(All models infer on NVIDIA RTX3070)

REFERENCES

[1] J. Peng, Y. Liu, S. Tang ... and Y. Ma, (2022). "PP-LiteSeg: A Superior Real-Time Semantic Segmentation Model." arXiv preprint arXiv:2204.02681.

[2] R. Gao, (2021). "Rethink dilated convolution for real-time semantic segmentation." arXiv preprint arXiv:2111.09957.

[3] I. Ulku, and E. Akagündüz, (2022). "A survey on deep learning-based architectures for semantic segmentation on 2d images. " Applied Artificial Intelligence, 1-45.

[4] J. Xu, Z. Xiong, and S. P. Bhattacharyya,(2022). "PIDNet: A Real-time Semantic Segmentation Network Inspired from PID Controller." arXiv preprint arXiv:2206.02066.

[5] J. Long, E. Shelhamer, and T. Darrell, (2015). "Fully convolutional networks for semantic segmentation." In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.

[6] Z. Kütük, and G. Algan, (2022). "Semantic Segmentation for Thermal Images: A Comparative Survey." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 286-295.

[7] C. Yu, J. Wang, C. Peng, C. Gao, G.Yu, and N. Sang, (2018). "BiSeNet: Bilateral segmentation network for real-time semantic segmentation." In: Proceedings of the European conference on computer vision ECCV, pp. 325-341.

[8] J. Yuan, Z. Deng, S. Wang, and Z. Luo (2020). “Multi receptive field network for semantic segmentation.” In: IEEE Winter Conference on Applications of Computer Vision WACV, pp. 1883-1892.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, (2017). “Attention is all you need.” Advances in neural information processing systems, 30.

[10] Z. Liu , Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, ... and B. Guo, (2021). “Swin transformer: Hierarchical vision transformer using shifted windows.” In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022.

[11] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, (2017). “Rethinking atrous convolution for semantic image segmentation.” arXiv preprint arXiv:1706.05587.

[12] J. Hu, L. Shen, and G. Sun,(2018). “Squeeze-and-excitation networks.” In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141.

[13] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, (2018). “Cbam: Convolutional block attention module.” In: Proceedings of the European conference on computer vision ECCV, pp. 3-19.

[14] C. ROTHER, V. KOLMOGOROV and A. BLAKE, (2004). “GrabCut" interactive foreground extraction using iterated graph cuts” ACM transactions on graphics TOG, 23(3): 309-314.

[15] L. Ding, , and G. Ardeshir, (2001). “On the Canny edge detector.” Pattern recognition, 34.3 721-725.

[16] I. Sutskever, and G. E. Hinton, (2012). “Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems, 25.