Construction Site Fence Recognition Method Based on Multi-Scale Attention Fusion ENet Segmentation Network

Xing Zhang
Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering
China Three Gorges University
Yichang, China
zhangxing@ctgu.edu.cn

Tinglong Tang*
Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering
China Three Gorges University
Yichang, China
tangtinglong@ctgu.edu.cn

Yirong Wu
Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering
China Three Gorges University
Yichang, China
yirongwu@hotmail.com

Tingwei Quan
Wuhan National Laboratory for Optelectronics
Huazhong University of Science and Technology
Wuhan, China
quantingwei@hust.edu.cn

***Abstract*—In this paper, we propose a fence recognition method based on the ENet (Efficient neural Network) segmentation network to address the problems of traditional segmentation networks, which have poor performance in recognizing fences with a large range of scale variations and hollow structures. Firstly, a multi-scale attention fusion ENet segmentation network is designed, which is trained using the fence with obvious color features. Then, a morphological algorithm is used to process the predicted image to restore the fence segmentation results. The designed multi-scale attention fusion segmentation network performs better on fence datasets than traditional methods. In addition, the activation function Leaky_Relu6 further enhances the stability and generalization ability of the network. The experiments are conducted on 540 fence images from different construction sites, and the computed IoU is 90%. The processing speed is about 28 frames per second. The experimental results show that our proposed network outperforms traditional segmentation algorithms in fence recognition performance, and achieves robustness in different construction scenarios while meeting the requirements of both accuracy and speed.**

Keywords- fence recognition; ENet; multi-scale attention; morphological algorithms;

I. INTRODUCTION

As a common safety protection facility on the construction site, the fence is usually used to isolate dangerous areas from the construction staff, which can play a certain protective role. However, when the staff is facing away from the fence, there will still be some safety hazards due to the blind spot. Therefore, this paper hopes to apply the method of deep learning to identify the fence on the site in real-time and accurately, and send a reminder when the construction staff and the fence are close, which can reduce the accident rate to a certain extent and ensure the safety of construction staff.

The construction site fence is mainly composed of thin red and white bars arranged in a certain shape, and its main shapes include vertical and cross-shaped patterns, as shown in Fig. 1. Traditional methods of acquiring segmented datasets involve pixel-by-pixel annotation of foreground pixels. However, pixel-by-pixel operations would require a lot of time due to the high density of the fence, and the hollow structure of the fence itself would contain a large amount of background information, making it difficult for the network to learn the correct features. Moreover, because the relative scales of the fence in different images can vary dramatically, the range of scale changes in the feature map is also large, which presents a challenge for traditional segmentation algorithms.

Figure 1. Example of fence.

As technology continues to evolve and replace older methods, there have been significant improvements in the accuracy and speed of multiscale object segmentation. However, not all segmentation algorithms can be applied to practical engineering projects. While accuracy and speed are important, other factors such as memory usage and model stability must also be considered. ENet (Efficient neural Network) [1] is a lightweight segmentation network that has a higher inference speed than traditional networks, making it more suitable for environments with limited memory, such as embedded devices.

Considering the structural characteristics of fence data, this paper proposes a fence recognition method based on improved ENet neural network. Firstly, a multi-scale attention fusion ENet neural network is constructed. Then, a new activation function is designed for the encoding part of the network to increase the model's stability and generalization ability. Finally, the network is trained using parts of the fence structure with clear color features, and the predicted image is processed using morphological algorithms to restore the entire fence structure. The proposed network has the following advantages:

DOI: 10.18293/SEKE23-081

- The multiscale attention fusion ENet neural network combines low-level positional information with high-level semantic information. The attention mechanism is used to capture the context dependencies of foreground pixels, allowing the model to obtain richer feature maps.
- A new activation function, Leaky_Relu6, is designed, which combines the advantages of Leaky_Relu [2] and Relu6 [3]. This function preserves useful information from the negative axis while suppressing the model's maximum output, preventing gradient explosion and improving model stability.
- Using morphological algorithms to recover the complete fence structure from partial fence structures does not require a large amount of memory, making it advantageous for the network to be used in environments with limited memory, such as embedded devices.

II. Related Work

A. Multi-scale Segmentation

In recent years, image segmentation has been widely used to identify target objects in images or videos and perform instance segmentation at the pixel level [4]. However, the segmentation of images with large-scale variations remains a challenging problem [5]. Olaf et al. proposed the U-Net [6] segmentation network, which uses an encoding-decoding network structure and skips connections in the encoding and decoding parts to achieve good performance in multi-scale medical image segmentation. Chen et al. developed the Deeplab [7] segmentation network, which uses dilated convolution and fully connected convolutional neural networks to further improve segmentation accuracy. Long et al. proposed the FCN [8] segmentation algorithm based on fully convolutional neural networks, which uses deconvolutional layers to output more refined image segmentation results and employs skip connections to increase the network's robustness. Zhang et al. employed a context encoding [9] network to obtain global information of the target and combined it with a common encoding-decoding network structure to output feature maps with richer multi-scale information.

B. Scene Segmentation

Yu et al. developed the BiSeNet [10] lightweight semantic segmentation network, which uses a bidirectional mechanism to acquire receptive fields that retain spatial and semantic information, allowing for faster network speeds without sacrificing accuracy. Mohan et al. proposed the EfficientPS [11] panoramic segmentation network, which can simultaneously segment background and foreground information. They introduced a new panoramic fusion module that dynamically adjusts the panoramic segmentation results based on the confidence levels of semantic and instance segmentation, greatly aiding autonomous driving environment perception. Liu et al. proposed the CRNet [12] segmentation network with a cross-reference mechanism that enhances model output feature representations by comparing similar features in two images, achieving small-sample image segmentation. Weng et al. introduced the DMA-Net [13] semantic segmentation network suitable for street view data, which aggregates feature maps generated by different convolutional layers through a multi-branch aggregation network to obtain multi-scale information of the target, achieving good performance on the CamVid dataset.

C. Segmentation with Attention Mechanism

Fu et al. proposed a dual attention mechanism [14] that combines channel and spatial attention to the aggregate context information and improves the expression ability of multi-scale feature maps, resulting in improved performance in scene segmentation. Hou et al. proposed the self-attention distillation network SAD [15], which optimizes low-level learning with high-level positional information and enhances high-level feature expression with low-level learned attention features, performing well in lane line segmentation. Tao et al. utilized a hierarchical multiscale attention mechanism [16] to predict different scale targets and output final results through pixel level operations, reducing memory usage by four times while improving accuracy and segmentation speed. Huang et al. developed CCNet [17] which uses a criss-cross attention module that cycles through the network to obtain horizontal and vertical contextual information of target pixels and generate richer feature maps.

III. Method

In this paper, we propose a fence recognition method based on an improved ENet neural network. Firstly, we construct a multi-scale attention fusion ENet neural network, and design a new activation function Leaky_Relu6 in the encoding part to replace PRelu [18], which improves the stability and generalization ability of the network. Then, we train the network using fences with obvious color features and restore the complete fence area from partially recognized fence structures through computer morphological algorithms.

A. ENet with Multi-Scale Attention Fusion

To effectively reduce the computational cost and time in segmentation tasks, this paper chooses the ENet segmentation network as the main component of the segmentation network. The ENet network structure is mainly composed of an initial module and several bottleneck modules, as shown in Fig. 2.

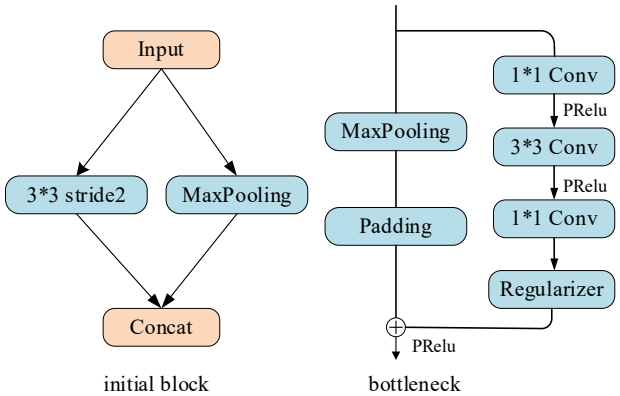


Figure 2. Main structure modules of ENet network.

In the initial block in Fig. 2, a 3x3 convolution with a stride of 2 and a maximum pooling are performed on the input. The results on both sides are merged into channels by concatenation. In the main branch of the bottleneck module in Fig. 2, the channel number is reduced by a 1x1 convolution, and a specific number of feature maps is output by a 1x1 convolution layer. If the bottleneck module is not down-sampled in the other branch, no operation is performed in this branch and it is directly added to the main branch. If it is down-sampled, this branch is first down-sampled by maximum pooling, then padded to achieve the same size as the feature map in the main branch.

The original ENet neural network consists of five parts, its structure is similar to an encoding-decoding structure, in which the first three parts extract feature information, and the last two parts are used to restore the feature map with the original size.

Generally speaking, for neural networks, the lower the feature map size and the less down-sampling operations, the richer the location information of small targets. The higher the feature map size and the larger the receptive field, the richer the semantic information. However, as the number of down-sampling operations increases, the location information of small targets becomes insufficient. Therefore, this paper combines the feature maps of the first and fourth parts of ENet, as well as the feature maps of the second and third parts, through a multi-scale fusion to combine the location information of the lower layers with the semantic information of the higher layers, thereby enhancing the model's feature learning ability.

At the same time, in each fusion process, in order to better aggregate foreground pixels and reduce the influence of interference pixels, Coordinate Attention (CA) [19] is added during the fusion process. CA is an attention mechanism that embeds positional information into channel attention by dividing channel attention into two dimensions of encoding. One dimension is used to obtain the dependency relationship between foreground pixels and contextual information, and the other dimension is used to preserve the positional information of foreground pixels, thereby enabling the feature map to have better direction and position expression ability.

B. Leaky_Relu6 Activation Function

In the encoding part of the original ENet network, PRelu was used as the activation function instead of Relu. Compared to the case where the negative output of Relu is constantly 0, PRelu can adjust the output value of the negative part adaptively by a learnable parameter. However, this occurs at the cost of the need of learning an extra parameter. Therefore, this paper leans towards the use of Leaky_Relu for the negative output, as shown in Fig. 3a. It outputs a small value to retain some useful information and avoid dead neurons without the need for additional parameter learning.

For the positive output, considering the high requirements for devices and networks on mobile devices, allowing unlimited output may lead to gradient explosion. Therefore, inspired by Relu6 activation function, as shown in Fig. 3b, a maximum value of 6 is set as the upper limit to suppress unlimited output. This can effectively solve the gradient explosion problem, prevent overfitting, and improve the model's generalization ability and stability.

Inspired by Leaky_Relu and Relu6 activation functions, this paper proposes a new activation function called Leaky_Relu6, as shown in Fig. 3c. It outputs an extremely small value when the input is less than 0. When the input is between 0 and 6, it outputs the same value as the input, and when the input is greater than 6, it outputs 6. This activation function is also used in the encoding component of the network, to replace the PRelu activation function in the original ENet network.

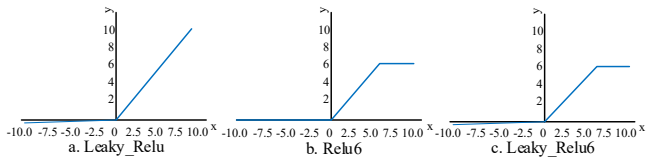


Figure 3. Three activation functions.

Combined with the above improvement points, the structure diagram of the network we proposed in this paper is shown in Fig. 4.

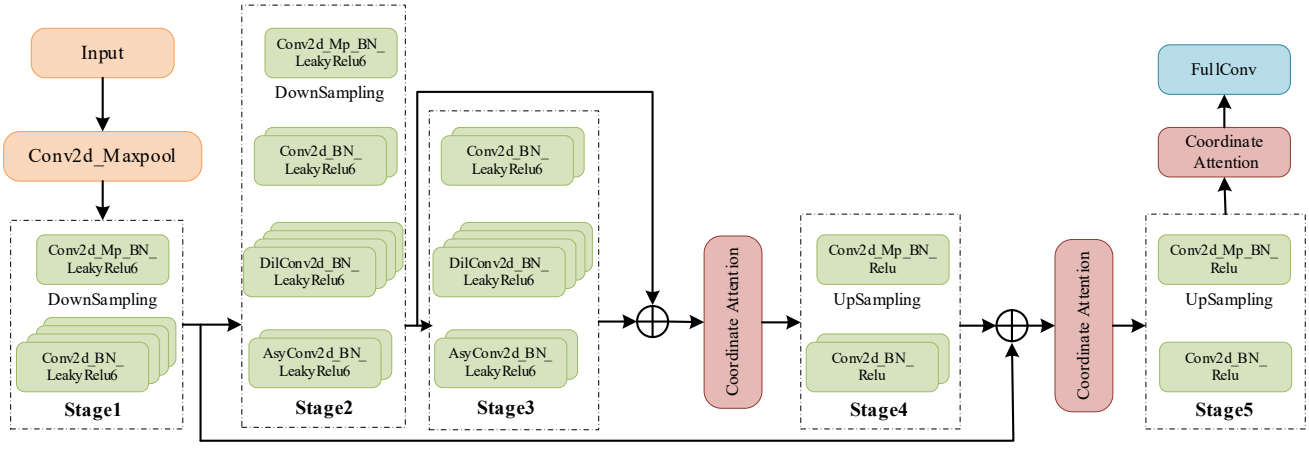


Figure 4. Structure diagram of the network we proposed in this article

C. Graphics Processing Algorithms

The image containing fence facilities is fed into the trained network to obtain the predicted image, which is then subjected to morphological operations such as closing, polygon dilation, and rectangle filling [20] to recover the complete fence structure from partial fence segmentation results. At the same time, the position corresponding to the fence area in the construction site picture is marked to realize the positioning of the fence area of the construction site

Firstly, a rectangular kernel of size 9*9 is set to divide the predicted image into several rectangular structural elements, which are then subjected to closing operation to fill the concave corners. The mathematical formula for the closing operation is as follows:

$$A \cdot B = (A \oplus B) \odot B \tag{1}$$

where A represents the predicted image, B represents the rectangular structural elements segmented by the rectangular kernel, and $\oplus$ represents the dilation operation. $\odot$ represents the corrosion operation. The formula for dilation operation is as follows:

$$P = (A \oplus B) = \{x, y | B_{x,y} \cap A \neq \emptyset\} \tag{2}$$

where $B_{x,y}$means that the origin of the rectangular structural element is moved to the point (x, y). The prediction image A is traversed by the structural element B, and if there is pixel intersection between B and A, the entire structural element B is retained. The corrosion operation is performed on the image P after the dilation operation, and the mathematical formula of the corrosion operation is as follows:

$$A \cdot B = P \odot B = \{x, y | B_{x,y} \in P\} \tag{3}$$

which means that structural element B goes through the binary image P after dilation operation. The pixels that intersect between structural element B and binary image P are retained, and other disjoint pixels are removed.

To prevent erroneous connections caused by closure operation, we set a vertical pixel threshold of 10. Columns with fewer than 10 pixels in the vertical direction are considered erroneous connections and are removed to prevent errors in subsequent steps. After polygon dilation, a rectangle filling operation is performed on the foreground pixels. Specifically, the number of pixels between the top and bottom pixels in each column are set to 255. Through these morphological image operations, the complete fence structure is successfully restored from the partially segmented fence structure. The implementation process of graphic processing algorithms is shown in Fig. 5.

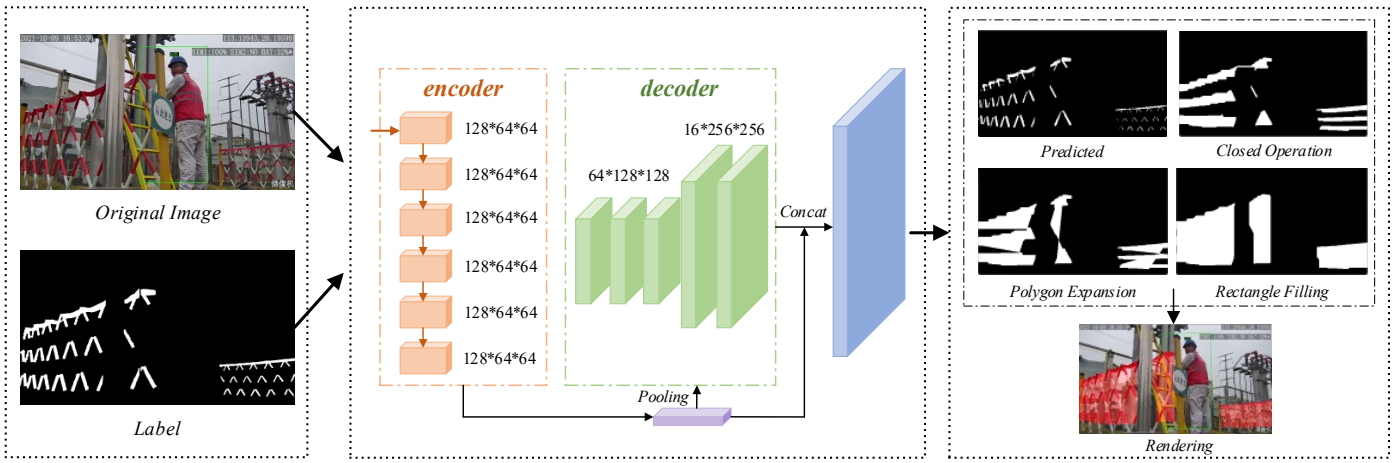


Figure 5. Flowchart of graphical processing algorithm

IV. Experiment Results and Analysis

The experiments are performed in Windows 10 system with a NVIDIA GTX1080 graphics card. The main algorithm in this article is implemented using Pycharm, where ENet training is carried out under the Pytorch framework 1.1.0. The libraries used in the implementation process mainly include Numpy, OpenCV, etc.

A. Experimental Data

The experimental data in this study was primarily obtained from the cameras located at power grid construction sites. Additionally, some common fence facility images are obtained through internet searches. The dataset contains 3980 fence images, including 240 samples with complex scenes, such as fences with inconspicuous color features or covered with plastic film, and approximately 300 negative samples, each with a size of 1280*720 pixels. Among them, the training set is divided into 2800 images, the validation set is divided into 640 images, and the remaining 540 images are used as the test set.

B. Evaluation Metrics

In this study, Mean Intersection-Over-Union (mIoU) [21], the time *t*(s) spent to process each image, and the image processing speed v (frames/s) are used as the performance evaluation metrics. The IoU is a standard measure for semantic

segmentation that calculates the ratio of the intersection to the union of two sets. The formulas for calculating the IoU and v are as follows:

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{4}$$

$$v = \frac{N}{\sum_i^N t_i} \tag{5}$$

In the above formulas, $target$ represents true pixel area of the safety fence and $prediction$ represents the pixel area that is actually predicted to be the security fence. N represents the total number of images, and t represents the time required to process each image.

C. *Analysis and Comparison of Results*

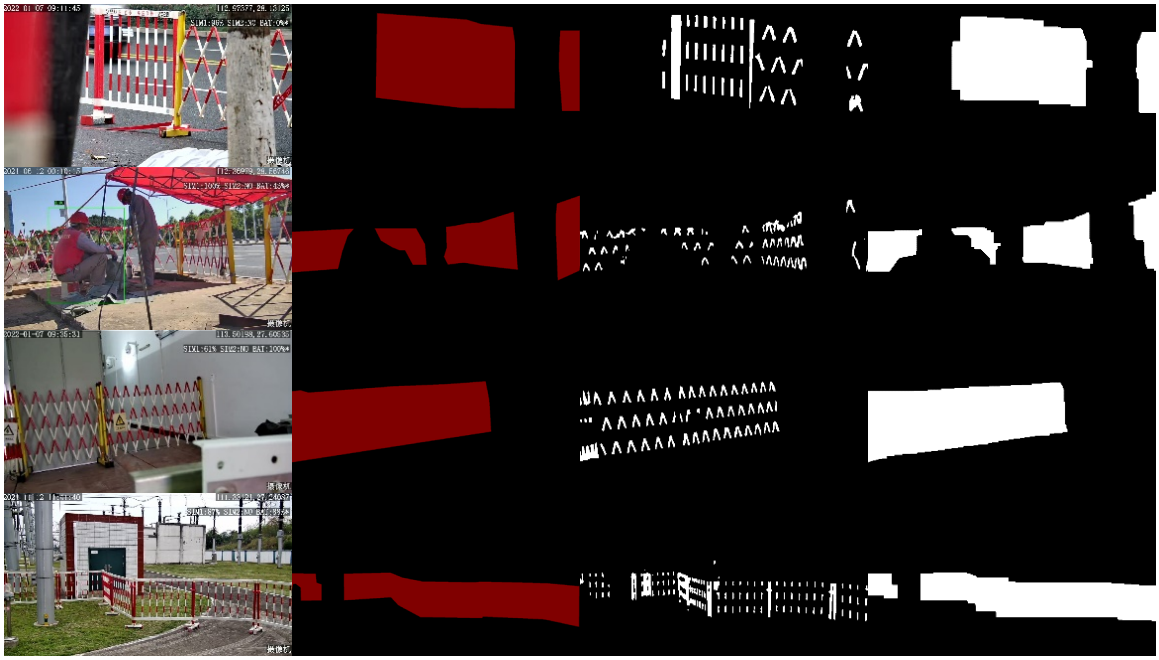
Firstly, based on the distribution characteristics of fence structures in construction site images, fences are classified into four types: vertical fences without occlusions, intersecting fences without occlusions, vertical fences with occlusions, and intersecting fences with occlusions. These four types are respectively denoted as type A, B, C, and D.

To test the effectiveness of the method proposed in this paper, the segmentation of fences of the above four types was tested individually (Fig. 6). Specifically, Fig. 6a displays the original images of fences captured in different construction scenarios and with different shapes, including a cross-shaped fence with obstructions in a road construction scene, a cross-shaped fence with obstructions in another road construction scene, a cross-shaped fence without obstructions in an indoor construction scene, and a vertical fence with obstructions in a substation construction scene. Fig. 6b shows the ground truth of the fences obtained using the labelme annotation software [22], Fig. 6c displays preliminary predicted fence regions generated using our improved ENet segmentation network, and Fig. 6d shows the fence regions recovered using graphical processing algorithms.

Table I shows the performance of the improved ENet network models and the original ENet network on the fence dataset. It can be observed that the improved network model achieves overall better recognition performance without sacrificing recognition speed, demonstrating the effectiveness of the proposed network architecture and new activation function.

Table II presents the results of our proposed method when four different types of fences are identified. From Table 2, it can be observed that our proposed method achieves an accuracy of around 90% for fence identification in different scenes. Especially for data with occlusions, it seems that our method performs well in identifying fence boundaries. Additionally, it seems that our proposed method meets the processing speed requirements while achieving high accuracy throughout the entire identification process.

Table III summarizes the segmentation accuracy of our proposed method for the 540 images in the four different fence types. It can be observed that our proposed method demonstrates good robustness in segmentation, achieving an overall recognition rate of approximately 90%. This may meet the requirements for fence recognition in construction sites and provides a solid foundation for future work.



a. origin picture b. ground truth c. predicted d. our method

Figure 6. Experimental results of the method in this paper

TABLE I. EXPERIMENTAL PERFORMANCE OF DIFFERENT MODELS

Model	mIoU	Average of speed(s)
ENet	0.854	0.036
ENet+Leaky_Relu6	0.863	0.035
Mul-scale ENet+CA+Leaky_Relu6	0.896	0.038

TABLE II. EXAMPLE OF FENCE RECOGNITION

Type	Origin Images	The Results of Our Method	IoU	Time
A			0.913	0.034s
B			0.887	0.037s
C			0.894	0.036s
D			0.885	0.039s

TABLE III. SEGMENTATION PERFORMANCE OF OUR METHOD

Type	Number of Tests	mIoU
A	116	0.925
B	147	0.894
C	143	0.912
D	134	0.883

V. CONCLUSIONS

This paper proposes a fence recognition method based on improved ENet neural network. The method utilizes the distinct color features of fences for neural network training and employs morphological algorithms to process the predicted images, enabling fast and accurate segmentation of the fence structures in construction site images. This method has the advantages of low cost, high accuracy, and low computational complexity. It can be used to develop safety guarantee systems for the stuff working at construction sites.

REFERENCES

[1] A. Paszke, A. Chaurasia, S. Kim, and E.Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," arXiv preprint arXiv:1606.02147, 2016.

[2] J. Xu, Z. Li, B. Du, and J. Liu, "Reluplex made more practical: Leaky ReLU," 2020 IEEE Symposium on Computers and Communications (ISCC). France, pp. 1-7, July 2020.

[3] Y. Zou, L. Zhao, S. Qin, M. Pan, and Z. Li, "Ship target detection and identification based on SSD_MobilenetV2," 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). China, pp. 1676-1680, June 2020.

[4] S. Minaee, Y. Boykov, F. Porikli, et al, "Image segmentation using deep learning: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 44, no. 7, pp. 3523-3542, February 2021.

[5] N. Xu, L. Yang, Y. Fan, et al, "Youtube-vos: A large-scale video object segmentation benchmark," arXiv preprint arXiv:1809.03327, 2018.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference. Germany, pp. 234-241, October 2015.

[7] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. 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, April 2017.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 3431-3440, June 2015.

[9] H. Zhang, K. Dana, J. Shi, et al, "Context encoding for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 7151-7160, June 2018.

[10] C. Yu, J. Wang, C. Peng, et al, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," Proceedings of the European Conference on Computer Vision (ECCV). pp. 325-341, September 2018.

[11] R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," International Journal of Computer Vision. vol. 129, no. 5, pp. 1551-1579, February 2021.

[12] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). USA, pp. 4165-4173, August 2020.

[13] X. Weng, Y. Yan, G. Dong, et al, "Deep multi-branch aggregation network for real-time semantic segmentation in street scenes," IEEE Transactions on Intelligent Transportation Systems. vol. 23, no. 10, pp. 17224-17240, February 2022.

[14] J. Fu, J. Liu, H. Tian, et al, "Dual attention network for scene segmentation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 3146-3154, June 2019.

[15] Y. Hou, Z. Ma, C. Liu, and C. Loy, "Learning lightweight lane detection cnns by self attention distillation," Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). pp. 1013-1021, October 2019.

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

[17] Z. Huang, X. Wang, L. Huang, et al, "Ccnet: Criss-cross attention for semantic segmentation," Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). pp. 603-612, October 2019.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," Proceedings of the IEEE International Conference on Computer Vision(ICCV). pp. 1026-1034, December 2015.

[19] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). USA, pp. 13713-13722, November 2021.

[20] S. Sattari and M. Izadi, "An improved upper bound on dilation of regular polygons," Computational Geometry. vol. 80, pp. 53-68, July 2019.

[21] M. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," International Symposium on Visual Computing. Springer, Cham, pp. 234-244, 2016.

[22] A. Torralba, B. Russell, and J. Yuen, "Labelme: Online image annotation and applications," Proceedings of the IEEE. vol. 98, no. 8, pp. 1467-1484, June 2010.