International Core Journal of Engineering Volume 7 Issue 7, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

A Brief Survey on Semantic Segmentation based on Deep Neural Network

Weina Zhou, Kun Chen

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, china.

Abstract

In recent years, semantic segmentation has become an important research topic in the field of machine vision. With the development of deep learning technology, image semantic segmentation based on deep neural networks has achieved remarkable results. Semantic segmentation is pixel-level image understanding, and each pixel in the image is assigned a category label. Semantic segmentation is widely applied in scenes such as autonomous driving, intelligent robots, human-computer interaction, etc. A large number of semantic segmentation methods have been proposed. In this paper, we introduce the background of semantic segmentation. Then, we divide the semantic segmentation methods based on deep learning into five categories and present the advantages and disadvantages of each class. Besides, we analyze publicly available datasets and evaluation metrics of semantic segmentation. Finally, this paper prospects the development trend of semantic segmentation.

Keywords

Semantic Segmentation; Deep Neural Network; Weakly Supervised Methods; RNN.

1. Introduction

Image semantic segmentation is a technology that enables the computer to automatically recognize the content of the image based on the semantic information. In the computer field, semantic segmentation of the image is pixel-level image understanding. Each pixel of the image is labeled with its category and assigned a predicted label. In recent years, semantic segmentation has attracted many scholars' attention, and its application fields are extensive. Such as self-driving cars, defect detection, and intelligent medical diagnosis, etc.

Before deep learning, traditional image segmentation algorithms divide the image into different regions based on the image's color, spatial structure, and texture. The traditional image segmentation algorithms contain the clustering segmentation method, threshold segmentation method, and image segmentation algorithm based on graph partitioning. Traditional image segmentation algorithms have low computational complexity, but segmentation performance is limited for image segmentation tasks in complex environments. [1] proposed a multi-layer neural network to automatically learn high-level features from a large amount of training data. Compared with the traditional segmentation algorithm, the semantic segmentation algorithm based on the deep neural network can obtain richer features. The semantic segmentation algorithm predicts each pixel's label by the extracted image features, thereby realizing the segmentation of different objects in the image. The performance of almost state-of-the-art public datasets is achieved through deep learning methods.

Some existing reviews have summarized the semantic segmentation technology [2-4]. [2] made a systematic summary and analysis of image semantic segmentation methods based on supervised learning. [3] introduced the traditional learning-based semantic segmentation in detail, such as the method based on support vector machine. [4] divided semantic segmentation methods into three

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

categories, namely methods based on manual engineering features, methods based on learning features, and methods based on weakly-supervised learning. According to the characteristics of the semantic segmentation model, this paper divides semantic segmentation methods into five categories and summarizes them. For each mentioned semantic segmentation model, the innovative work is introduced in detail. Besides, this paper introduces the classic datasets and objective evaluation indicators of semantic segmentation. Finally, we present several valuable research points in this field.

2. Deep learning architectures for semantic segmentation

[5] presented fully convolutional networks (FCN), which realize the pixel-level classification. Compared with the traditional convolutional neural network, FCN achieved better performance on semantic segmentation. In recent years, many semantic segmentation algorithms have been developed from fully convolutional neural networks. We have categorized the semantic segmentation models into five different categories: methods based on decoder structure, methods based on context information, methods based on GAN, methods based on recurrent neural network (RNN), and methods based on weak supervision. Besides, we summarize the innovative work and existing problems of the proposed semantic segmentation model.

2.1 Decoder-structure-based methods

Ronneberger et al. [6] proposed an encoder-decoder model called U-Net, which alleviates the problem of reduced resolution of feature maps due to down-sampling operations. U-Net adopts a U-shaped architecture and is mainly used in medical image segmentation. U-Net gradually restores the object's details and the resolution of the feature map through skip connection and up-sampling, effectively fusing low-level detail information and deep semantic information to achieve a good segmentation effect. At the same time, U-Net can perform accurate segmentation by a small number of training samples.

To solve the problem of road scene segmentation, Badrinarayanan et al. [7] proposed SegNet, which comprises an encoder network and a corresponding decoder network. As shown in Figure 1, the encoder network in SegNet is based on VGG [8]. The decoder consists of the deconvolution layer and the upsampling layer. The decoder upsamples its input using the pooling indices of its corresponding encoder layer to generate sparse feature maps. Simultaneously, SegNet adds a batch normalization layer, which speeds up the convergence speed of the network and suppresses overfitting. SegNet has low computational complexity and is very efficient in terms of memory and computational time.

In [9], the authors proposed Deeplabv3+, a network based on the encoder-decoder architecture. The proposed decoder fuses deep semantic features and shallow detail features to refine the segmentation results. The authors also applied the depthwise separable convolution to the entire model, which improves the calculation speed. Deeplabv3+ reaches 89.0% mIOU on the PASCAL VOC 2012 datasets, achieving a balance between the accuracy and speed of image semantic segmentation.

The method based on decoder structure enhances the output features of the entire network by fusing detail features and semantic features, thereby improving the performance of the model.

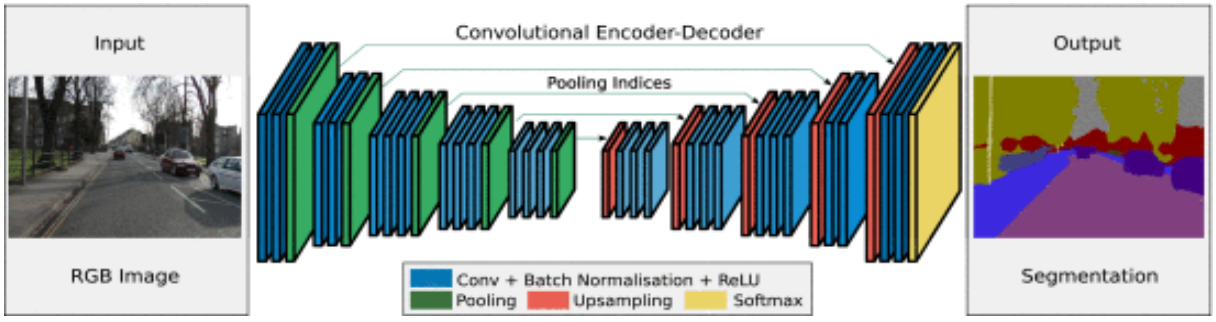

Figure 1. The architecture of SegNet

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

2.2 Context-based methods

Generally, the context information can perceive semantics. Therefore, the introduction of context information can improve the accuracy of semantic segmentation. [10] proposed ParseNet, an end-to-end semantic segmentation network. ParseNet obtains the global context information of the image through the global average pooling, helping to classify local confusion. Experimental results show that ParseNet reaches 69.8% mIOU on the PASCAL VOC 2012 datasets with small additional computational overhead. Similar to ParseNet, PSPNet [11] also introduces global context information. As shown in Figure 2, to realize the ability of global context. the authors proposed a pyramid pooling module to aggregate contextual information based on different regions. Besides, the auxiliary loss function is added in the base network training process, which reduces the optimization difficulty. And the entire network performs well in complex scene analysis tasks.

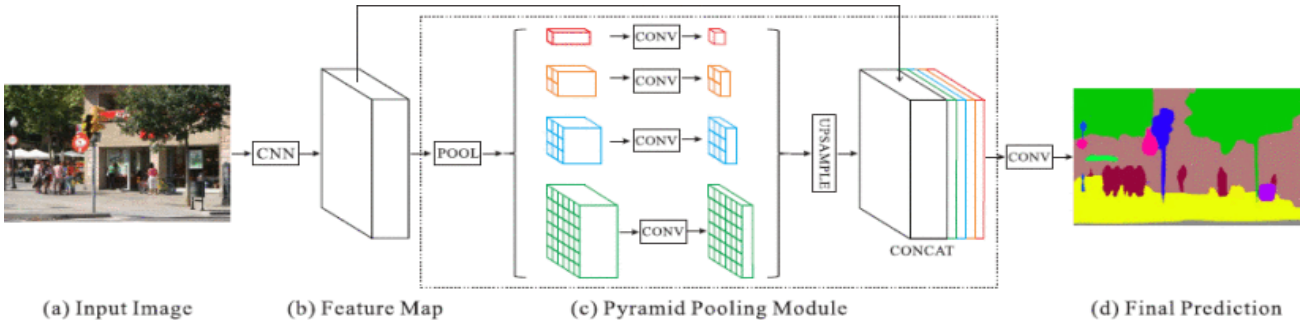

Figure 2. The architecture of PSPNet

In [12], Hang Zhang et al. proposed a context encoder module, which integrates an encoding layer. The context encoder module can capture context information and highlight critical feature mapping. The authors introduced semantic encoding loss to standardize training, which strengthens the network learning of semantic context. Besides, the proposed EncNet improves the segmentation performance of small objects and achieves 85.9% mIOU on the PASCAL VOC 2012 datasets. To solve the problem of scale variety of street scenes, [13] proposed DenseASPP, which can encode multi-scale context information while maintaining a large enough receptive field. DenseASPP performs advanced performance in street scene segmentation tasks.

The context-based methods extract dense image features by introducing global information or multi-scale context information. These methods make full use of features at different locations to improve segmentation accuracy.

2.3 GAN-based methods

In recent years, the generative adversarial network (GAN) [14] has attracted the attention of researchers. GAN consists of two parts: a generator and a discriminator. The generator and discriminator conduct iterative adversarial training to improve the performance of the network gradually. Inspired by the generative adversarial network, LUC et al. [15] introduced GAN to the field of semantic segmentation for the first time. The network contains a segmentation network(generator) and an adversarial network(discriminator). The discriminator distinguishes the images generated by the generator and the ground truth label maps. Combining multi-class cross-entropy loss and adversarial terms, the authors proposed a new objective function to train the entire network.

Compared with traditional GAN, [16] proposed SegAN, an end-to-end adversarial neural network. The authors proposed multi-scale L1 loss, which considers different levels of features and optimizes segmentation network in backpropagation. SegAN has better performance for image segmentation problems and is suitable for medical image segmentation. To solve the inconsistency of low-level local features and the inconsistency of high-level semantic features in the pixel-level classification.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

[17] proposed the Macro-Micro Adversarial Network (MMAN). It is worth noting that MMAN has two discriminators: Macro D and Micro D. The former is used to enhance semantic consistency. The latter is used to improve local consistency. Besides, MMAN alleviates poor convergence in the processing of high-resolution images by the adversarial network.

The GAN model has the ability to generate data and continuously optimize the generated data, which is the key to the problem of small sample feature learning. However, the optimization of the GAN model is unstable, and its performance needs to be improved.

2.4 RNN-based methods

The recurrent neural network (RNN) is another main-stream model of deep learning. RNN has the characteristic of memorizing historical information and can use the information of the previous moment to guide the output of the next moment. Applying RNN in image segmentation can fully consider the correlation between image pixels.

Inspired by RNN, Francesco et al. [18] proposed the ReSeg model, whose structure is shown in Figure 3. A pre-trained VGG-16 network extracts the features, and the ReNet [19] layers capture context dependencies from the extracted features. ReNet is an improved RNN model that can effectively acquire the image's global features and context information by scanning the image horizontally and vertically. Experiment results show that ReSeg can efficiently handle image segmentation tasks. However, some categories are not well segmented.

Compared with RNN, Long Short Term Memory (LSTM) architecture can store and retrieve information quickly or for a long time. [20] proposed an image segmentation method based on LSTM-RNN, which is based entirely on learning. LSTM-RNN does not contain any additional post-processing steps and is well adapted to complex image scenes with low computational complexity. [21] proposed the Graph-LSTM structure, extending sequence data to graph structure data. Graph-LSTM uses superpixels as graph nodes and adaptively constructs an undirected graph topology. The undirected graph topology is based on superpixels and corresponding spatial connections. This method reduces the cost of redundant calculations and enhances feature representation.

RNN and the convolutional layer can be combined into a deep neural network. The convolutional layer has good performance in extracting the local spatial features. RNN can capture correlation features between pixels of images.

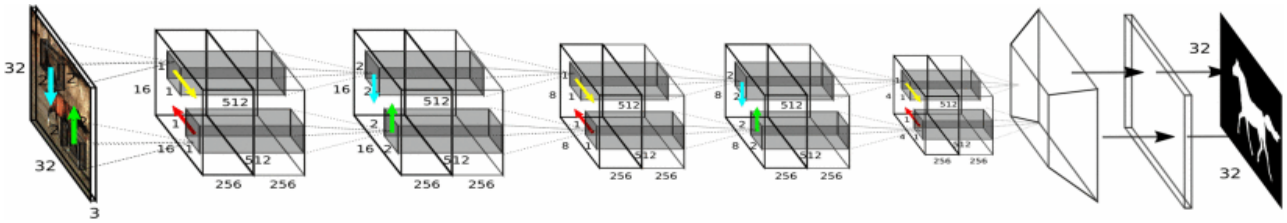

Figure 3. The architecture of ReSeg.

2.5 Weakly supervised methods

The semantic segmentation model based on deep learning has continuously made breakthroughs in segmentation effects in recent years. However, as the depth and width of the network continue to increase, this requires a larger dataset size, and collecting labeled data is a time-consuming and labor-intensive task. Therefore, researchers turn to the semantic segmentation method of weakly supervised learning. [22] proposed a method based on bounding box labeling to provide labeling information. The proposed BoxSup method takes the image marked by the bounding box as the training sample. It gradually restores the segmentation mask by iterating between automatically generating region proposal and training the convolutional network. MANINIS et al. [23] proposed a semi-automatic segmentation CNN framework called DEXTR. The extreme points (leftmost, rightmost, bottom, and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

top pixels) of the object are used as the network's input to realize semantic segmentation. By creating a heatmap with the activated regions of extreme points. The heatmap is cascaded with the RGB channels of the image to form a four-channel input feature. This method is used for segmentation experiments on five different datasets, and all have achieved good performance.

3. Datasets and metrics

With the development of semantic segmentation technology, more and more algorithms are proposed. Large-scale datasets and unified evaluation indicators help to evaluate these algorithms systematically. Therefore, this section mainly introduces public datasets and evaluation indicators commonly used in image semantic segmentation experiments.

3.1 Datasets

3.1.1 PASCAL VOC 2012

PASCAL VOC 2012 [24] contains training set, validation set, and test set. The corresponding number of images are 1464, 1449, and 1452 respectively. There are 21 categories consists of animals, transportation, indoor objects, and background. The dataset is publicly available and is the evaluation standard for the challenge of object segmentation, recognition, and detection in the field of computer vision.

3.1.2 Cityscapes

Cityscapes [25] is a large pixel-level annotated dataset collected from street scenes in 50 different cities. Cityscapes consists of 7 target categories: ground, architecture, road signs, nature, sky, people, and vehicles. The numbers of the training set, validation set, and test set in this dataset are 2975, 500, and 1525 respectively. In addition, to evaluate the performance of the classification network based on weakly supervised learning, Cityscapes also provides 20,000 coarsely segmented images.

3.1.3 PASCAL Context

PASCAL Context [26] is extended on the PASCAL VOC 2010 dataset. The dataset has a total of 540 categories, of which 59 semantic categories are frequently used. PASCAL Context contains 10103 images for training and validation sets, and 9637 images for the testing set.

3.1.4 CamVid

CamVid [27] is the earliest semantic segmentation dataset applied in the field of autonomous driving. The dataset contains five different video sequences, and 700 frames are manually annotated by the annotation software, and the resolution size of each image is 960×720. CamVid includes 32 categories, such as buildings, walls, trees, sidewalks, and traffic lights.

3.1.5 KITTI

KITTI [28] is a dataset used to evaluate 3D target detection and tracking performance. The dataset is often used in autonomous driving scenes and consists of more than 200,000 images with 3D annotated targets. KITTI consists of 11 categories in total, including scene data such as urban areas, rural areas, and highways.

3.1.6 MS COCO

MS COCO [29] is a large-scale target detection and semantic segmentation dataset. The dataset has a wide variety of targets, and the images are mainly collected from indoor and outdoor scenes. In MS COCO, there are 165482,81208, and 81434 images for training, validation, and testing, respectively. The dataset contains 91 categories and is often used for image recognition and semantic segmentation tasks.

3.2 Evaluation Metrics

Running speed, model complexity, and accuracy are used to evaluate the performance of semantic segmentation algorithms. Frame Per Second (FPS) is used to measure the running speed of the algorithm. Model complexity is measured by model parameters and Floating Point Operations

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

(FLOPs). In this section, we mainly introduce the accuracy of the algorithm. We assume that the total number of categories is c+1, which includes the background category. p_{ij} represents the number of pixels in which the ith category is predicted to be the jth category, p_{ii} represents the number of positive samples that are predicted to be correct. p_{ji} represents the number of pixels in the jth class that is predicted to be the ith class.

The metric of pixel accuracy (PA) means the ratio of the number of correctly classified pixels to the total number of pixels in the image.PA is defined as:

$$PA = \frac{\sum_{i=0}^{c} p_{ii}}{\sum_{i=0}^{c} \sum_{j=0}^{c} p_{ij}} \tag{1}$$

Intersection over union (IoU) presents the ratio of the intersection of the predicted map and the true annotated map to the union of these two maps.

$$IoU = \frac{\sum_{i=0}^{c} p_{ii}}{\sum_{i=0}^{c} \sum_{j=0}^{c} p_{ij} + \sum_{i=0}^{c} \sum_{j=0}^{c} p_{ji} - \sum_{i=0}^{c} p_{ii}} \tag{2}$$

Mean intersection over union (mIoU) denotes the average value of the accumulated IoU values of each class of image.

$$mIoU = \frac{1}{c+1} \sum_{i=0}^{c} \frac{p_{ii}}{\sum_{j=0}^{c} p_{ij} + \sum_{j=0}^{c} p_{ji} - p_{ii}} \tag{3}$$

4. Future Directions

4.1 Real-time semantic segmentation

In recent years, with the development of autonomous driving, people have higher and higher requirements for real-time semantic segmentation. At present, a large number of real-time semantic segmentation methods have been proposed. For example, [30] proposed DFANet, a lightweight real-time semantic segmentation network. Experiments on the Cityscapes dataset prove that DFANet achieves considerable accuracy with small model complexity and fast running speed. However, there is still a lot of space for improvement. In the future, semantic segmentation will explore how to improve the running speed while maintaining high precision.

4.2 Video semantic segmentation

At present, the main direction of semantic segmentation is concentrated at the single image level. However, applications in areas such as intelligent transportation and intelligent surveillance are all based on video sequences. The main difficulty of video semantic segmentation is the spatio-temporal information in the feature map. Although researchers have proposed some semantic segmentation methods for video sequences, they cannot meet the development needs of video semantic segmentation. Compared with image semantic segmentation, video semantic segmentation has greater practical significance.

5. Conclusion

In this paper, we have provided a brief survey of semantic segmentation based on deep neural networks. According to the algorithm's characteristics and the model's structure, the investigated methods have been divided into five categories, such as the method based on the decoder structure. We introduce the main ideas and advantages of each method in detail. Six publicly available datasets and have been reported and described. Besides, we analyze standard metrics used for semantic segmentation. Finally, we have also discussed some potential research directions for semantic segmentation, such as real-time semantic segmentation and video semantic segmentation.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 52071200, 61404083)

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

References

[1] G. E. Hinton and R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, vol. 313, pp. 504 - 507, 2006.

[2] A. Garcia-Garcia, S. Orts, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Rodríguez, "A survey on deep learning techniques for image and video semantic segmentation," Appl. Soft Comput., vol. 70, pp. 41-65, 2018.

[3] M. Thoma, "A Survey of Semantic Segmentation," ArXiv, vol. abs/1602.06541, 2016.

[4] H. Yu et al., "Methods and datasets on semantic segmentation: A review," Neurocomputing, vol. 304, pp. 82-103, 2018.

[5] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 640-651, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," ArXiv, vol. abs/1505.04597, 2015.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 2481-2495, 2017.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR, vol. abs/1409.1556, 2015.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," ArXiv, vol. abs/1802.02611, 2018.

[10] W. Liu, A. Rabinovich, and A. Berg, "ParseNet: Looking Wider to See Better," ArXiv, vol. abs/ 1506. 04579, 2015.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230-6239, 2017.

[12] H. Zhang et al., "Context Encoding for Semantic Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7151-7160, 2018.

[13] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for Semantic Segmentation in Street Scenes," 2018 IEEE/ CVF Conference on Computer Vision and Pattern Recognition, pp. 3684-3692, 2018.

[14] I. Goodfellow et al., "Generative Adversarial Networks," ArXiv, vol. abs/1406.2661, 2014.

[15] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic Segmentation using Adversarial Networks," ArXiv, vol. abs/1611.08408, 2016.

[16] Y. Xue, T. Xu, H. Zhang, L. Long, and X. Huang, "SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation," Neuroinformatics, vol. 16, pp. 383-392, 2018.

[17] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-Micro Adversarial Network for Human Parsing," in ECCV, 2018.

[18] F. Visin et al., "ReSeg: A Recurrent Neural Network-Based Model for Semantic Segmentation," 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 426-433, 2016.

[19] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. C. Courville, and Y. Bengio, "ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks," ArXiv, vol. abs/1505.00393, 2015.

[20] W. Byeon, T. Breuel, F. Raue, and M. Liwicki, "Scene labeling with LSTM recurrent neural networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3547-3555, 2015.

[21] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic Object Parsing with Graph LSTM," in ECCV, 2016.

[22] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1635-1643, 2015.

[23] K. Maninis, S. Caelles, J. Pont-Tuset, and L. Gool, "Deep Extreme Cut: From Extreme Points to Object Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 616-625, 2018.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0054

[24] M. Everingham, S. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," International Journal of Computer Vision, vol.111, pp. 98-136, 2014.

[25] M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213-3223, 2016.

[26] R. Mottaghi et al., "The Role of Context for Object Detection and Semantic Segmentation in the Wild," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 891-898, 2014.

[27] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," Pattern Recognit. Lett., vol. 30, pp. 88-97, 2009.

[28] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," The International Journal of Robotics Research, vol. 32, pp. 1231 - 1237, 2013.

[29] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in ECCV, 2014.

[30] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9514-9523, 2019.