Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 7438914, 11 pages
https://doi.org/10.1155/2020/7438914

Research Article

Semantic Segmentation Algorithm Based on Attention Mechanism and Transfer Learning

Jianfeng Ye,[1,2] **Chong Lu**,[3] **Junfeng Xiong**,[1] **and Huaming Wang**[1]

[1]*Nanjing University of Aeronautics and Astronautics, Nanjing, China*
[2]*Xinjiang Vocational and Technical College of Communications, Urumqi, China*
[3]*Xinjiang University of Finance and Economics, Urumqi, China*

Correspondence should be addressed to Jianfeng Ye; y50345279@outlook.com

Received 9 May 2020; Revised 23 July 2020; Accepted 29 July 2020; Published 19 August 2020

Academic Editor: Giuseppe D'Aniello

In this paper, we propose a semantic segmentation algorithm (RoadNet) for auxiliary edge detection tasks with an attention mechanism. RoadNet improves the dispersion of the low-level features of the network model and further enhances the performance and applicability of the semantic segmentation algorithm. In RoadNet, a fully convolutional neural network is used as the basic model, an auxiliary loss in the image classification, multitask learning in machine learning, and attention mechanism in natural language processing. To improve the generalization of the model, we select and analyze a proper domain difference measure. Subsequently, the context semantic distribution module and the annotation distribution loss are designed based on the context semantic encoding structure. The domain discriminator based on the adversarial training and the adversarial training algorithm based on transfer learning are then well integrated to provide a transfer learning-based semantic segmentation algorithm (TransRoadNet). The experimental results indicate that the proposed TransRoadNet and RoadNet overperform their equivalent comparison models.

1. Introduction

Application of deep learning methods in image classification achieves remarkable results; see, e.g., [1–4]. Deep learning is also extensively applied in image semantic segmentation. For instance, E. Shelhamer et al. [5] present a fully convolutional neural network (FCN) for image segmentation. The combination of transfer learning and deep learning [6] is also used to introduce the concepts and methods of deep learning in a variety of research fields in social and natural sciences [7–9]. However, in practical applications, the efficacy of deep learning based methods is challenged by the availability of enriched datasets, their inference accuracy, and generalization performance of the deep learning models. In this paper, we address these three challenges:

(1) Although PASCAL VOC2012 [10], CIFAR-10/100 [11], and Cityscapes dataset [12] are able to provide powerful training data sources, multi-view observation scenes are further required to be constructed for the complex urban road images. In this paper, using Eagle Eye, the high-altitude road monitoring dataset is formed, and the virtual images are collected via the communications between the graphics library and the monitoring game to make the virtual dataset. Hereafter, we simply refer to these enriched datasets as Surv-Citispace and Virt-Citispace.

(2) Regarding the attention mechanism and to make the low-level features of the network, the auxiliary task branch for edge detection is designed based on objects' shape and edge information. Meanwhile, an auxiliary task learning module and an attention-constant residual network are constructed to form a semantic segmentation model, namely, RoadNet. In order to improve the receptive field of the semantic segmentation task, global pooling concepts and comprehensive cascading ideas are utilized to further improve the atrous spatial pyramid pooling and design a cascaded atrous spatial pyramid pooling.

(3) We further investigate the transfer learning algorithm for RoadNet from the perspectives of domain difference measurement, semantic distribution loss, and adversarial learning and then design a semantic segmentation model, namely, TransRoadNet, based on transfer learning. TransRoadNet effectively reduces the performance loss of basic model, RoadNet, in the process of migration and deployment on different data (i.e., Cityspace to Virt-Citispace and/or Surv-Citispace to Cityspace).

2. Related Works

2.1. Attention Mechanism. Chen et al. [13] introduce using conditional random fields to the FCN as a post-processing algorithm. Zhao et al. [14] also design a pyramid pooling module to aggregate the context information of different regions by combining four feature maps of different scales to improve the capability obtaining global information of the neural networks.

Regional-Convolutional Neural Network (R-CNN) in [15] triggers the application of target detection convolutional neural networks based on the candidate regions. He et al. [16] suggest using shared convolutions to speed up the calculation of R-CNN. The region-of-interest pooling layer is also designed by Girshick [17] based on a spatial pyramid pooling which is able to pool the considered regions with different sizes, into a fixed-size feature vector. Ren et al. [18] suggest handing over the task of finding candidate target areas to a deep convolutional neural network and propose the Region Proposal Network (RPN). Further, a network branch is added by He et al. [19] based on RPN to predict the segmentation mask of the target object. They further expand their method from the original simple target recognition to instance segmentation. The current research results are greatly influenced by related thoughts [20, 21].

In the above works, the attention mechanism is often utilized to explicitly model the interdependence between the semantic features of $F(x_t, W_t)$ and x_t. This is done through combining the attention residual module (ARM) with the residual module and self-attention mechanisms. Due to adaptive enhancement of the channel graph of relevant semantics, it is therefore possible to replace the feature fusion in the original residual network and further enhance the ability to express the relevant semantics of the residual module.

2.2. Receptive Fields and Auxiliary Tasks. From the multitasking perspective, Badrinarayanan et al. [22] introduce the encoder-decoder structure into the FCN, where the pooling layer index is retained to store more image information in the encoding stage. In this stage, the pooling layer index is used to restore image loss information.

Holmstrom [23] indicated that occasional addition of noise during the training can enhance the generalization capability of the network model. In contrast to other methods which are focused on enhancing the training effect of the auxiliary tasks, RoadNet is focused on enhancing the training effect of the main task. In this context, the edge detection is often considered as an auxiliary task. The low-level shared network mainly considers the edge and shape information of the object; hence, it can obtain more features regarding the differences in the object categories. The annotated images which are required for edge detection can be simply attained from the semantically segmented annotated images.

Regarding the receptive field, Fisherand Koltun [24] showed that the FCN upsampling is unable to restore the information lost. This is because of the pooling layer downsampling without loss. To address this issue, they suggest atrous convolution, where the original convolution range is extended thus increasing the receptive field of the network. ASPP is also applied by several researchers; see, e.g., [25].

Here, we combine the cascading idea in DenseASPP with the global pooling branch in ASPPv2 and propose the cascade atrous spatial pyramid pooling (CascadeASPP). In our proposed design, the atrous convolution of multiple atrous rates is connected step by step. It provides a larger receptive field, improves the pixel sampling density of the atrous convolution, and hence forms more receptive fields to provide a higher level of size invariance. Moreover, to tackle the degradation problem of atrous convolution, here the global context information is obtained through the global pooling branch.

2.3. Transfer Learning and GAN. To find a suitable difference domain measure, we train the following three methods on the basic FCN network and RoadNet with the above-mentioned three transferred datasets:

(1) Correlation Alignment (CORAL) proposed by Bao Sun et al., which is an unsupervised domain adaptive algorithm [26]

(2) Maximum Mean Discrepancy (MMD) as one of the most commonly used distance measures in transfer learning [27]

(3) Contrastive Domain Discrepancy (CDD) which adds category information based on MMD and hence measures the intraclass and interclass differences across domains [28]

The best representation is discovered by the feature-representation-based transfer through feature transformation. The context semantic encoding [29] (CSE) captures the global context scene information and improves the scene-related feature map. Nevertheless, the context semantic encoding only predicts the existence of the category as prior knowledge of the scene with obvious defects. Hence, a semantic distribution loss is proposed to replace the semantic encoding loss. Particularly, in the proposed semantic distribution loss, the ability of the model to predict the existence of the categories and the proportion of the categories in the image is essential, adding more prior knowledge of scenes and the relationship between categories to the model.

The generative adversarial network is a network model proposed by Goodfellow et al. [30]. It can better grasp the

global information by discriminating against the network compared to the direct use of the loss function. Moreover, TransRoadNet integrates GAN's domain adversarial ideas and replaces the image generation network in GAN with source and target domain feature extractors to extract the image features. The task of discriminating the network in GAN is to determine the extraction of the image features from the source or target domain images. The domain-invariant features are extracted by the encoder as much as possible so that the discriminator cannot distinguish between the two domains. Meanwhile, the discriminator needs to distinguish the two domains as much as possible to conduct adversarial training.

3. Network Structure

3.1. Attention Residual Module. The original residual module is shown in Figure 1(a) as

$$\begin{aligned} y_1 &= F(x_l, W_l), \\ x_{l+1} &= h(x_l) + f(y_l), \end{aligned} \tag{1}$$

where x_l and x_{l+1} represent the input and output of the l-th layer, respectively, F shows the residual function, h denotes the identity mapping function, and f is the rectified linear activation function. Although the identity mapping function in the residual module can ensure no loss in the information flow, the information flow of the entire network includes loss due to the activation function. Therefore, f also becomes an identity mapping function to obtain an enhanced residual module, namely, the identity residual module [31], ensuring the flow of the information between the layers without a loss (Figure 1(b)).

The mathematical expression is as follows:

$$\begin{aligned} x_{l+1} &= x_l + F(x_l, W_l), \\ x_L &= x_1 + \sum_{i=1}^{L-1} F(x_i, W_i). \end{aligned} \tag{2}$$

Based on the backpropagation chain rule, the following partial derivative is obtained:

$$\begin{aligned} \frac{\partial \text{Loss}}{\partial x_1} &= \frac{\partial \varepsilon}{\partial x_L}\frac{\partial x_L}{\partial x_1}, \\ &= \frac{\partial \varepsilon}{\partial x_L}\left(1 + \frac{\partial \sum_{i=1}^{L-1} F(x_i, W_i)}{\partial x_L}\right). \end{aligned} \tag{3}$$

Equation (3) indicates that the loss gradient can be transferred to any residual module without loss. Even the loss gradient of any residual module can be converted without loss to the remaining residual modules; hence, the probability of vanishing the gradient is reduced.

Nevertheless, if each channel of the feature graph is assumed to be the semantic feature response graph of the segmentation target, there must be a correlation between the corresponding graphs of the semantic features of various segmentation targets in the image. The semantic features of x_l and y_l in the residual module are not consistent and are not added directly. Hence, the self-attention mechanism is inserted into the fusion of x_l and y_l in the identity residual module to explicitly model the interdependence between semantic features. Using the interdependence between the channels, it is possible to improve the interdependent features as well as the representation of the specific semantic features:

$$\begin{aligned} y_l &= F(x_l, W_l), \\ x_{l+1} &= x_l + y_l P(x_l, y_l). \end{aligned} \tag{4}$$

The input feature map of the attention residual module (Figure 2) is $X \in R^{C\times H\times W}$. A novel feature map, $Y \in R^{C\times H\times W}$, is then obtained, after two rounds of batch normalization, convolution, and activation function. Hence, X and Y are reorganized into $X' \in R^{C\times N}$ and $Y' \in R^{C\times N}$, respectively. Matrix multiplication is also performed on the transpose of X', and Y'. After normalizing the exponential function, the channel attention graph $A \in R^{C\times H\times W}$ is finally obtained:

$$a_{i,j} = \frac{\exp(x_i, y_j)}{\sum_{i=1}^{C} \exp(x_i, y_j)}, \tag{5}$$

where $a_{i,j}$ represents the influence factor of the i-th channel of X to the j-th channel of Y. Matrix multiplication is conducted on A and Y', and $E \in R^{C\times H\times W}$ is readjusted as the improved feature map. The ultimate output feature map, $O \in R^{C\times H\times W}$, is then attained by adding elements of E and X.

3.2. RoadNet and Auxiliary Edge Detection Tasks. Figure 3 represents the RoadNet structure. In particular, the training task signal of the auxiliary task has specific domain information to improve the main task generalization effect. Following the pyramid network structure of FPN [32, 33], a semantic segmentation network model is designed to test the auxiliary tasks, including a top-down basic network, a horizontal connection, and a bottom-up edge detection auxiliary network. The accurate edge detail information is then obtained from shallow features, and then the semantic information is attained from the deep features. Consequently, the lack of image detail information in the original semantic segmentation network is eliminated.

The network takes an image of any size as input and then calculates a feature map of multiple scaling ratios using the basic network. The network is also divided into five stages based on the size of the feature map. The relative scale of the feature map output by the last residual module to the input image in each stage is 4, 8, 16, and 32, respectively.

By upsampling of the image of the high-level feature pyramid, the edge detection auxiliary network restores its resolution. The basic network is also connected with the edge detection auxiliary network through horizontal connections to merge the feature maps of the same size. Furthermore, using the Canny algorithm, the annotated image of the edge detection auxiliary network is obtained from the annotated image of the semantic segmentation [34].

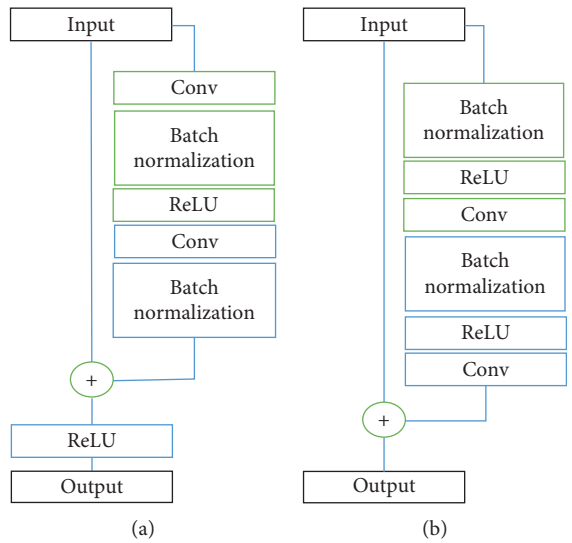


Figure 1: The residual module. (a) The original residual module. (b) The identical residual module.

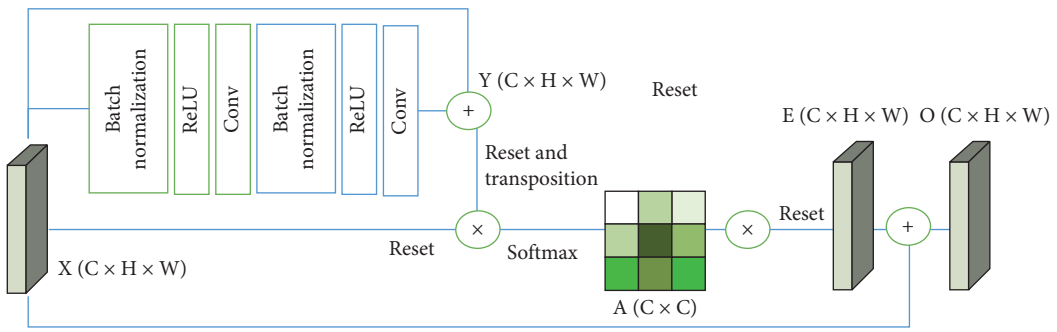


Figure 2: The attention residual module (ARM).

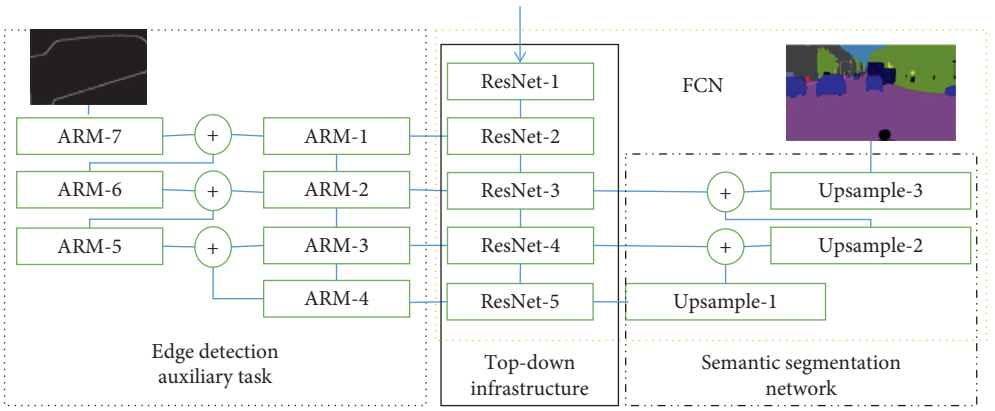


Figure 3: The RoadNet.

The loss function of the edge detection network takes multi-class empirical cross-entropy to normalize the predicted feature map exponentially. The calculation formula is

$$D_{i,j,n} = \frac{e^{X_{i,j,n}}}{\sum_{k=0}^{C} e^{X_{i,j,k}}}. \tag{6}$$

Then, the cross-entropy is calculated as

$$p(x, y) = \begin{cases} x, & \text{if } x = y, \\ 1 - x, & \text{otherwise.} \end{cases}$$

$$\text{Loss} = \frac{1}{MN} \sum_{i=0}^{M} \sum_{j=0}^{N} \left(- \sum_{k=0}^{C} p\left(\mathrm{D}_{i,j,k}, \mathrm{Y}_{i,j}\right) \log\left(p\left(\mathrm{D}_{i,j,k}, \mathrm{Y}_{i,j}\right)\right)\right), \tag{7}$$

where $Y_{i,j}$ is pixel i, j in the image, $D_{i,j,n}$ represents pixel i, j after the exponential normalization of the n-th channel of the image, $X_{i,j,n}$ is pixel i, j of the n-th channel of the image, M is the image length, N is the image width, and C is the category number.

3.3. CascadeASPP. ASPP has gained a large receptive field; however, a huge deal of image information is lost within the calculation process due to the low pixel sampling rate. For example, the receptive field size is 13 for a 3*3 atrous convolution with an atrous rate of 6; however, only 9 pixels are sampled for calculation. Then, the pixel sampling rate is 0.05. By connecting two convolutions with an atrous rate of 3 in series, the receptive field size is also 13, while 25 pixels are sampled for calculation. The pixel sampling rate is 0.15, which is more than three times the pixel sampling rate of the former. By a higher atrous rate, this effect becomes more obvious which is overcome by the proposed model effectively.

The global pooling branches and all atrous convolutions are cascaded through CascadeASPP (Figure 4). After 1×1 convolution and batch normalization, it is then upsampled to the preferred spatial dimension. For feature fusion, it is then merged with other atrous convolutions with different atrous rates. Through the cascading between different sizes of atrous rates, 13 sizes of receptive fields are covered. In the meantime, the coverage and atrous convolution pixels are sampled with a higher density.

3.4. Transfer Learning Mechanism. Using the feature-representation transferring technique, the difference between the target domain and the source domain is added to the loss function of the network model. Thus, the difference between the target domain and the features of the source domain is minimized through model training. After comparative testing, the MMD difference measure is selected as the loss function to design a context semantic distribution (CSD) module. The structure is illustrated in Figure 5. It is observed that the input feature map of the context semantic distribution module passes through two fully connected layers. The proportion of categories in the scene becomes an output of the fully connected layer, i.e., category distribution information. Consequently, for this category of distribution information as well as for the annotated image, the semantic distribution loss is calculated. The other fully connected layer outputs the scaling factor of the input feature map and then multiplies the input feature map and the scaling factor and by the channel as the output of the module. It is aimed at strengthening the feature maps related to the current scene based on the prior knowledge of the scene and also weakening the feature maps which are not related to the current scene. The category distribution information for the model inference graph is then determined, and the semantic distribution loss is calculated within the category distribution information of the annotated images.

The semantic distribution information of the annotated image is denoted as the feature vector p with the length of C. In this model, each value indicates the ratio of the pixels occupied by the category c in the annotated image to the total pixels of the image (i.e., the percentage of the image occupied by each category). The calculation formula is stated as

$$\mathbf{P}_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} f\left(\mathbf{Y}_{i,j}, c\right), \tag{8}$$

where

$$f(x, c) = \begin{cases} 1, & \text{if } x = c, \\ 0, & \text{othrewise.} \end{cases} \tag{9}$$

The semantic distribution information of the inference graph is then determined as

$$\widehat{p}_c = \frac{1}{HW} \sum_{i=1}^{W} \sum_{j=1}^{H} \frac{\mathbf{Y}'_{i,j,c}}{\max\left(\mathbf{Y}'_{i,j}\right)}. \tag{10}$$

where C is the total number of categories in the source domain dataset, Y is the annotated image, Y' shows the model inference graph, H is the pixel height of the annotated image, and W is pixel width of the annotated image.

Using a multi-class cross-entropy loss function, the semantic distribution loss is determined for the semantic distribution information of the annotated image and the inference graph, which is

$$\text{Loss} = - \sum_{c=0}^{C} p_t \log\left(\widehat{p}_t\right). \tag{11}$$

3.5. TransRoadNet. To simplify the training process, a gradient inversion layer is added between the domain discriminator and the basic network as a connection layer. However, the gradient inversion layer is corresponding to the identity mapping over the forward propagation. It includes no other operations and the input is directly outputted to the next layer. During the back propagation, the gradient inversion layer obtains the gradient from the next layer multiplied by −1, before passing to the previous layer. The structure of TransRoadNet is illustrated in Figure 6.

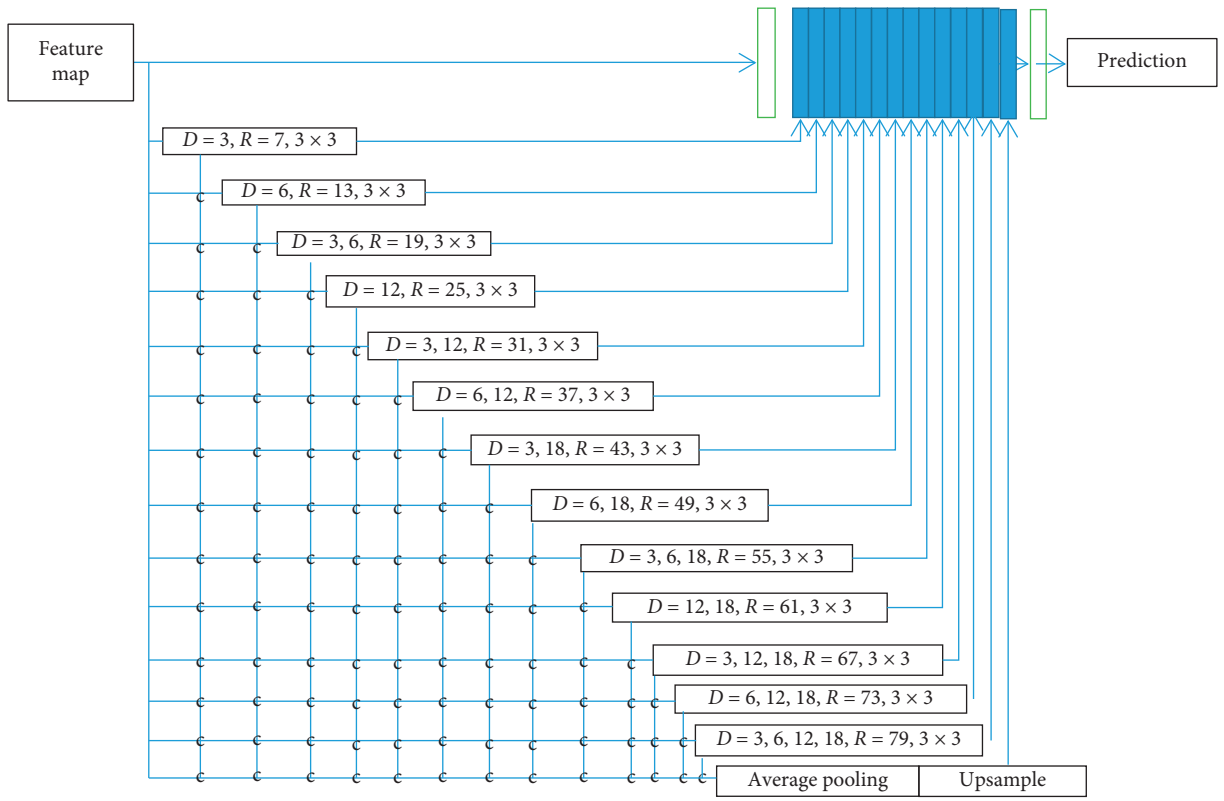


Figure 4: The CascadeASPP.

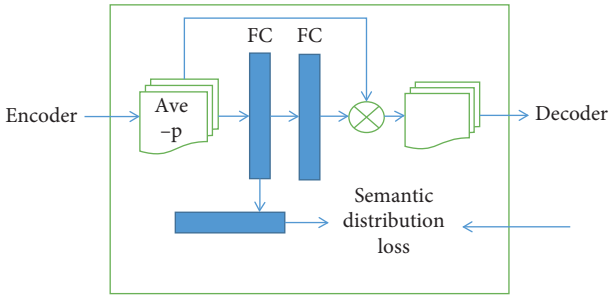


Figure 5: The context semantic distribution.

4. Experiments

4.1. Datasets. Based on the Cityscapes dataset [12], to collect the urban road traffic images, the surveillance video by Eagle Eye camera is considered as the source. An image from the video is intercepted by the dataset at given intervals and a total of 400 images are collected. Subsequently, the dataset is divided into a test set including 200 images and training set including 200 images (i.e., ratio of 5 : 5). This dataset is referred to as Surv-Cityscapes.

Grand Theft Auto V (GTA5) is selected by the virtual dataset as the virtual data collection virtual environment. GTA5 is started from Render Doc with a resolution of 1920×1080. The character is manipulated to drive the vehicle and to select the first angle of view. In total, 4000 images are collected, and they are randomly classified into 2000 test set images and 2000 training set images with a ratio 5 : 5; this dataset is referred to as Virt-Cityscapes.

Principal component dithering and random image cropping along with other algorithms are also utilized to augment the dataset and to create transfer learning datasets on these three city image datasets.

Three public datasets were also used in the experiment, including PASCAL VOC2012 [10] and CIfar-10 and CIfar-100 [11]. The CIFAR-10 dataset consists of 60,000 32×32 color images from 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images in the dataset. Cifar-100 dataset includes 100 classes, each containing 600 images. Each category includes 500 training images and 100 test images. The dataset, PASCAL VOC2012, supports image recognition tasks such as classification, target detection, and semantic segmentation. In our experiment, we use the semantic segmentation sub-dataset of PASCAL VOC2012.

4.2. Attention Residual Module Testing. For the first residual module in the attention residual network, the channel correlation between $F(x_1, W_1)$ and x_1 is obtained. The channel correlation heatmap is visualized in Figure 7. This figure illustrates the correlation between any channels of the two feature maps represented by the color of the corresponding cell. A higher (lower) correlation is shown by a lighter (darker) color. As it is seen in Figure 7, the color of the correlation heat map becomes significantly lighter after the attention mechanism. This means that the correlation in the feature map channel is significantly enhanced; therefore, the correlation between features is improved by the attention residual module.

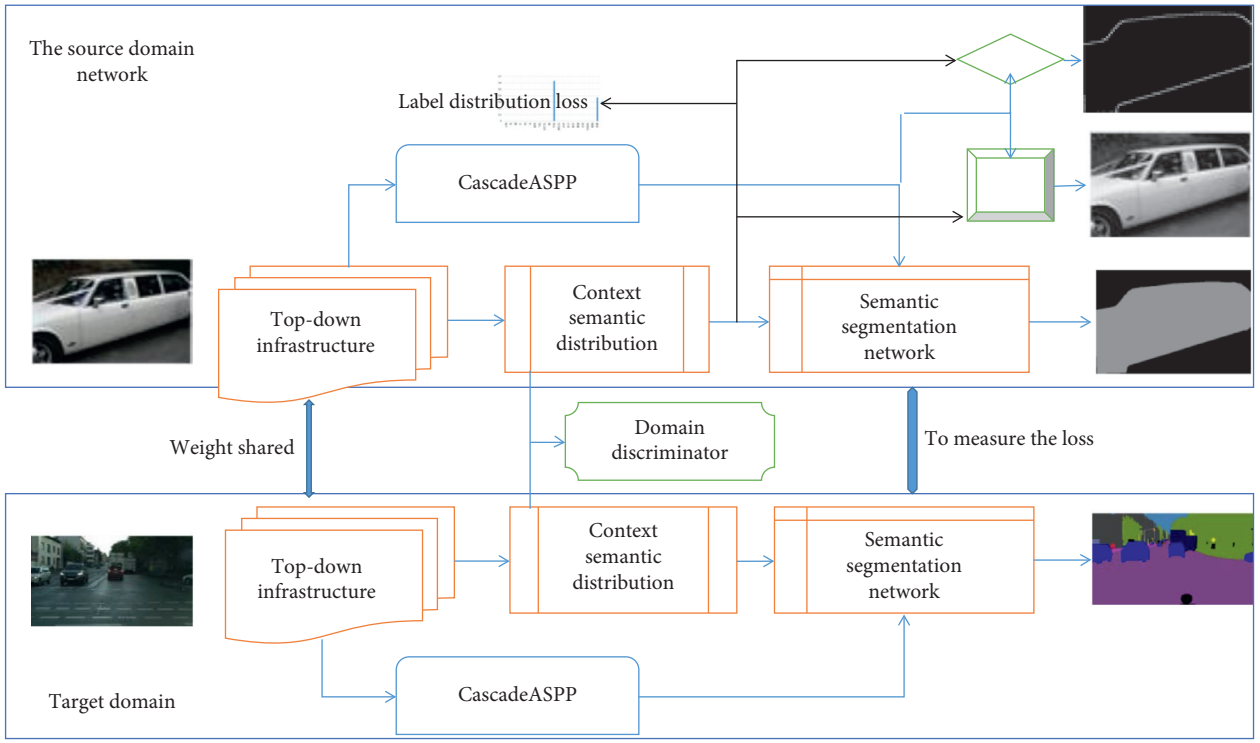
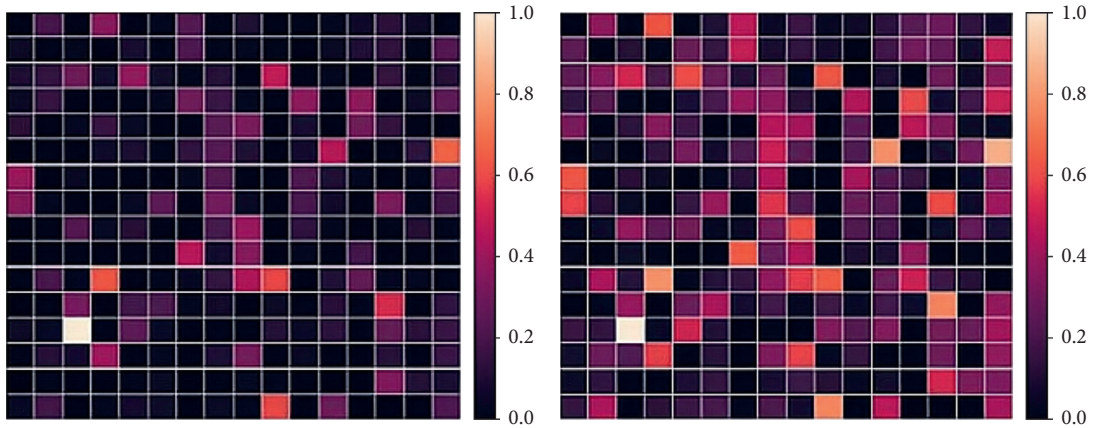


Figure 6: The TransRoadNet.



(a) (b)

Figure 7: The channel correlation heatmap. (a) Before the attentional mechanism. (b) After the attentional mechanism.

To obtain ResNet, the original residual modules are stacked, and the identity residual modules are also stacked to obtain IdentityResNet. To obtain AttentionResNet, the attention residual modules are stacked. Using Xavier algorithm [35], the weights are initialized, and the ResNet model which is trained on the ImageNet dataset is used as the pretraining model. The test results are presented in Table 1.

Compared to the original residual network, in CIFAR-100 dataset, the attention residual network in the 50-layer network is 1.45% lower. Furthermore, compared to the original residual network, the attention residual network in the 101-layer network is 1.49% lower. It is also seen that, using the attention residual module, the probability of convergence problems and degradation problems is greatly reduced.

4.3. CascadeASPP Testing. Here, we compare the FCN basic model, FCN-ASPP, and FCN-CascadeASPP on multiple datasets. The results of these comparisons are shown in Table 2. As it is seen, the model evaluation metric for the ASPP structure is greatly enhanced in comparison with the basic model in all datasets, which are also further enhanced by using CascadeASPP. These results suggest that the context

Table 1: The comparison of the residual network performance.

Model	Error rate (%)
ResNet-50 [4]	27.26
IdentityResNet-50 [14]	26.43
AttentionResNet-50	25.81
ResNet-101	26.51
IdentityResNet-101	25.98
AttentionResNet-101	25.02

Table 2: The comparison of CascadeASPP performance.

Test dataset	Model	PA (%)	MIoU (%)	FWIoU (%)
VOC2012	FCN	90.30	65.30	85.00
	FCN-ASPP	94.10	78.50	90.80
	FCN-CascadeASPP	94.80	80.20	91.30
Cityscapes	FCN	88.60	53.60	82.10
	FCN-ASPP	92.50	70.30	88.60
	FCN-CascadeASPP	93.20	72.60	89.80
Surv-Cityscapes	FCN	88.60	38.50	86.50
	FCN-ASPP	98.10	46.10	96.90
	FCN-CascadeASPP	98.10	47.50	97.10

mechanism is an important factor and a larger receptive field is essential for capturing further contextual information and prior knowledge of the scene.

4.4. CSD Testing. The encoder in RoadNet is the top-down basic network, while the bottom-up semantic segmentation main network is the decoder. The context semantic encoding module and the context semantic distribution module are, respectively, added to the FCN and RoadNet, and the new models are referred to as FCN-CSD, FCN-CSE, RoadNet-CSD, and RoadNet-CSE. We examine these models on three transfer learning datasets. The test results shown in Table 3 suggest the following:

(1) The context semantic distribution module has only 0.2% and 0.4% performance improvement on Surv-Cityscapes transfer learning dataset and around 3% performance improvement on both Cityscapes and Virt-Cityscapes transferred dataset. The reason is the fixed recording position and angle of the Surv-Cityscapes surveillance camera. This results in a fixed image scene, and hence its prior knowledge of the scene is relatively simple. Nevertheless, the model performance is improved by the context semantic distribution module as it adds more scene prior information to the model.

(2) For the three transfer learning datasets and the proposed network model, a performance improvement of about 0.2% to 3% is achieved by adding the context semantic distribution module and transferring the model. The results also indicate a further performance improvement in the context semantic encoding module. This validates the effectiveness of the context semantic distribution module.

Table 3: The comparison of the target domain MIoU.

Model	Cityscapes migrated dataset (%)	Surv-Cityscapes migrated dataset (%)	Virt-Cityscapes migrated dataset (%)
FCN	53.20	17.50	21.10
FCN-CSE	54.70	17.80	22.80
FCN-CSD	55.10	17.90	23.40
RoadNet	58.60	23.40	24.10
RoadNet-CSE	59.60	23.60	26.30
RoadNet-CSD	60.20	23.60	26.90

4.5. RoadNet Testing. To obtain a larger receptive field, CascadeASPP is added to the jump connection of RoadNet. To compare with EncNet, the number of training epochs is consistent and set to 62500. Table 4 represents the test results. Compared with the basic model, FCN is improved by 18.9%, 29.3%, and 12.8%, respectively. Moreover, it is enhanced by 1.7%, 2.1%, and 2.2% compared to EncNet's semantic segmentation model.

4.6. TransRoadNet Testing. To validate the model, the semantic segmentation network model, TransRoadNet, based on transfer learning is examined on the three transfer learning datasets. The training parameters are consistent with that of RoadNet, for which the results are recorded in Table 5. The average merge ratio of TransRoadNet is 62.7%, 30.6%, and 35.8%, respectively, which is 4.1%, 24.4%, and 12.7% higher than the model without transfer learning and 1.9%, 3.7%, and 4.4% higher compared to the common transfer learning algorithm.

Table 4: The comparison of the performance of the semantic segmentation network model.

MIoU	RoadNet (%)	FCN (%)	DeepLabv3 [36] (%)	ExFuse [37] (%)	DFN [38] (%)	DANet [39] (%)	EncNet [40] (%)
VOC2012	84.60	65.70	80.50	80.80	82.70	82.50	82.90
Cityscapes	82.90	53.60	79.50	80.10	80.30	80.50	80.80

Table 5: The comparison of the performance of semantic segmentation models based on transfer learning.

Dataset	Model	Before the migration MIoU (%)	After the migration MIoU (%)
Cityscapes	FCN wild	56.20	60.50
Migrated dataset	Curriculum DA	56.30	60.80
	TransRoadNet	58.60	62.70
Surv-Cityscapes	FCN wild	5.90	24.30
Migrated dataset	Curriculum DA	6.10	26.90
	TransRoadNet	6.20	30.60
Virt-Cityscapes	FCN wild	21.10	27.10
Migrated dataset	Curriculum DA	22.30	31.40
	TransRoadNet	23.10	35.80

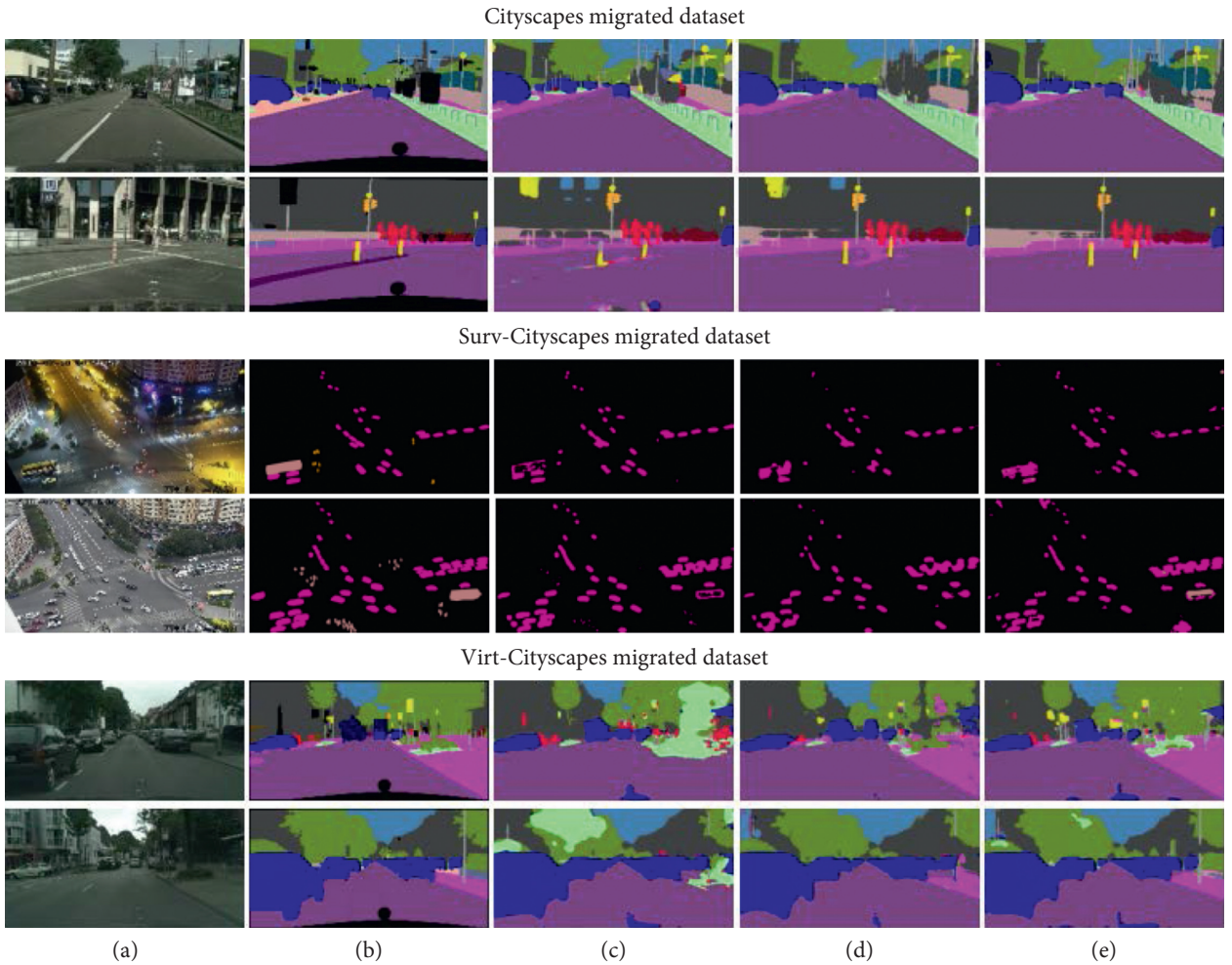


Figure 8: The comparison of semantic segmentation model inference based on transfer learning. (a) Artwork. (b) Mark figure. (c) FCN Wild. (d) Curriculum DA. (e) TransRoadNet.

For Cityscapes transfer learning dataset, the performance of the transfer learning algorithm is only enhanced by about 4%. This is owing to the limitations of the urban transferred dataset itself. The deviation of the dataset and the performance loss are small; hence, the effect of the transfer learning algorithm is not as clear as it is in the remaining datasets.

Figure 8 represents the inference graph of the semantic segmentation model based on transfer learning, where graph (a) shows the original image, graph (b) represents the annotation image, and graph (c), graph (d), and graph (e) denote the inference graphs of the semantic segmentation model based on transfer learning. It is observed that TransRoadNet is noticeably better than other semantic segmentation models in terms of transfer learning in edge segmentation effect and target classification accuracy.

5. Conclusion

Based on Cityscapes, two datasets with various perspectives of urban roads and their transferred datasets are constructed. RoadNet designed based on ARM and CascadeASPP possesses good portability and performance. TransRoadNet based on CSD shows a higher performance in the experiments compared to the un-transferred RoadNet and the transfer learning algorithms.

Data Availability

All data included in this study are available upon request to the corresponding author (e-mail address: hmwang@nuaa.edu.cn).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under grant no. 61363066.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Neural Information Processing Systems*, pp. 1106–1114, Lake Tahoe, NV, USA, December 2012.

[2] K. S. A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, pp. 1–14, San Diego, CA, USA, May 2015.

[3] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, October 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

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

[6] D. George, H. Shen, and E. Huerta, "Deep transfer learning: A new deep learning glitch classification method for advanced LIGO," 2017, https://arXiv:1706.07446.

[7] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proceedings of the Conference on Neural Information Processing Systems*, pp. 136–144, Barcelona, Spain, December 2016.

[8] Y. Xu, S. J. Pan, H. Xiong et al., "A unified framework for metric transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1158–1171, 2017.

[9] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 7167–7176, Honolulu, HI, USA, July 2017.

[10] M. Everingham, S. M. A. Eslami, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[11] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[12] M. Cordts, M. Omran, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Las Vegas, NV, USA, June 2016.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, Honolulu, HI, USA, July 2017.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2014.

[17] R. Girshick, "Fast R-CNN," in *International Conference on Computer Vision*, pp. 1440–1448, Las Condes, Chile, December 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, pp. 2980–2988, 2018.

[20] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 6819–6829, Seoul, Republic of Korea, November 2019.

[21] D. Marin, Z. He, P. Vajda et al., "Efficient segmentation: learning downsampling near semantic boundaries," in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 2131–2141, Seoul, Republic of Korea, November 2019.

[22] 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, no. 12, pp. 2481–2495, 2017.

[23] B. Holmstrom, "Moral hazard in teams," *The Bell Journal of Economics*, vol. 13, no. 2, pp. 324–340, 1982.

[24] Y. Fisher and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016.

[25] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[26] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 2058–2065, Phoenix, AZ, USA, February 2016.

[27] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by Kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[28] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 4893–4902, Long Beach, CA, USA, June 2019.

[29] H. Zhang, K. J. Dana, J. Shi et al., "Context encoding for semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 7151–7160, Salt Lake UT, USA, 2018.

[30] I. Goodfellow, J. Pougetabadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of thee Neural Information Processing Systems*, pp. 2672–2680, Montreal, Canada, December 2014.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision*, pp. 630–645, Amsterdam, The Netherlands, October 2016.

[32] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 936–944, Honolulu, HI, USA, July 2017.

[33] X. Liang, H. Zhou, and E. Xing, "Dynamic-structured semantic propagation network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 752~761, Salt Lake, UT, USA, 2018.

[34] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 249–256, Sardinia, Italy, May 2010.

[36] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution or semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, pp. 833–851, Munich, Germany, September 2018.

[37] Z. Zhang, X. Zhang, C. Peng, C. Peng, Xi. Xue, and J. Sun, "ExFuse: enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, pp. 273–288, Munich, Germany, September 2018.

[38] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 1857–1866, Salt Lake, UT, USA, June 2018.

[39] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3146–3154, Long Beach, CA, USA, June 2019.

[40] H. Zhang, K. J. Dana, J. Shi et al., "Context encoding for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, Salt Lake, UT, USA, June 2018.