Automation and Machine Learning (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2023.040105
ISSN 2516-5003 Vol. 4 Num. 1

Two-step Domain Adaptive Semantic Segmentation Algorithm of Driving Scene

Zhou Su[1,2,a], Yuan Tanghu[1,b,*], Yi Yuqian[2,c]

[1]School of Automotive Studies, Tongji University, Shanghai, China
[2]Sino-German College, Tongji University, Shanghai, China
[a]zhousu@tongji.edu.cn, [b]2368073615@qq.com, [c]yuqian_yi@163.com
[*]Corresponding author

Keywords: Automated driving, Computer vision, Semantic segmentation, Antagonistic

Abstract: Semantic segmentation of driving scenes is an important task in the field of automatic driving perception. The existing semantic segmentation based on deep learning method needs a lot of manpower cost to label data, and the changeable driving scenes will also lead to the performance degradation of semantic segmentation model in practical application. Therefore, in this paper, labeled computer-generated images are used as source domain data, unlabeled real driving scene images are used as target domain data, and unsupervised domain adaptive method is used to transfer knowledge when the target domain data is unlabeled. Based on the one-step domain adaptive semantic segmentation algorithm, a two-step domain adaptive semantic segmentation algorithm is proposed, and an image quality evaluation module is introduced to optimize the algorithm, so as to improve the performance index of cross-domain semantic segmentation tasks to some extent and reduce the dependence on data labeling and labeling cost.

1. Introduction

The vision tasks of autonomous vehicles mainly include classification, target detection, image segmentation, depth estimation, target tracking and so on. On the one hand, the current domain adaptation work is relatively applied to classification tasks, and most of these domain adaptation work is carried out in handwritten numeral classification or office31 article classification tasks; In addition to classification tasks, domain adaptive methods can also be applied to face recognition [1], pedestrian re-recognition [2-4], object detection [5-12], semantic segmentation [13] and so on, but there are relatively few studies at present. On the other hand, the methods of semi-supervised domain adaptation, unsupervised domain adaptation [13], including self-supervised domain adaptation are also research hotspots.

Aiming at the problem of strong data dependence and high labeling cost in semantic segmentation of driving scenes, this paper adopts unsupervised one-step domain adaptive method based on antagonistic learning to realize cross-domain segmentation, and then proposes a two-step domain adaptive method based on image transformation, and introduces an image quality evaluation algorithm to optimize the two-step domain adaptive segmentation method.

2. Single-step Domain Adaptive Network Framework

Traditional image segmentation methods are based on threshold, histogram and region growth, k-means clustering, etc. However, with the development of deep learning technology, image segmentation models based on deep learning methods are used to solve segmentation problems, such as FCN model, which is to integrate existing classification networks such as VGG16 and Google Net.

However, the image segmentation model based on deep learning has an obvious disadvantage, that is, when the domain adaptive method is not used, the inconsistent data distribution between the training set and the test set will lead to the decline of the cross-domain segmentation performance of the model. In order to solve this problem, the current mainstream solution is based on confrontation training, that is, the domain invariant features are extracted through the game between the generator and the discriminator, so that a model that has migrated from the source domain but still performs well in the target domain is trained without using the target domain label. This model consists of two parts, one is the partition network, and the other is the discriminator network.

The framework of one-step domain adaptive semantic segmentation network based on antagonistic learning consists of segmentation network and discriminator network, in which the segmentation network is used to predict the segmentation result and the discriminator network is used to judge whether its input comes from the target domain or the semantic segmentation output from the source domain.

The segmentation network is a Deeplabv2 segmentation model with Resnet101 as the backbone network, using spatial pyramid pooling structure (ASPP) as the final classifier, and finally applying an up-sampling layer to match the size of the input image.

The discriminator network is composed of five convolution layers. The convolution kernel has a size of 4×4 and a step size of 2. The number of channels in these five convolution layers is 64, 128, 256, 512 and 1 respectively.

3. Two-step Domain Adaptive Semantic Segmentation Algorithm

The single-stage domain adaptive model extracts the domain-invariant expression features, but it can't capture the pixel-level and low-level domain movement and can't be visualized.

In order to solve this problem, a two-step domain adaptive method based on image generation method can be adopted. This method is divided into two stages. The first stage is to transform the image in the source domain into the style of the target domain by using the unsupervised image transformation model. The second stage is to take the sample whose style has been transformed in the first step as the intermediate domain and align its features with the target domain. This stage adopts the single-step domain adaptive method.

However, the image generated by the image transformation model may be distorted and blurred, so in order to ensure the quality of the generated image, a scoring algorithm for image quality evaluation is introduced, and sample weights based on sample quality are introduced in the second stage of domain adaptive training, thus optimizing the two-step domain adaptive semantic segmentation algorithm.

3.1. Unsupervised Image Transformation Model

In order to reduce the domain offset distance, the two-step domain adaptive segmentation algorithm uses an unsupervised image transformation framework to generate the intermediate domain. The framework structure consists of three groups of convolutional neural network modules, namely encoders E1 and E2, decoders G1 and G2, and discriminators D1 and D2. The encoder and decoder are combined to form an automatic variational encoder structure, and the decoder and discriminator are combined to form a generation countermeasure network structure.

The structure of automatic variational encoder (VAE) is shown in Figure (1). Before generating the code, the encoder outputs two vectors M and σ with the same dimensions. Vector σ represents the variance of noise and is a parameter learned from the network. Vector e obeys normal distribution, and vector c is the generated code, which is obtained by adding noise to vector m. The function of VAE is to reconstruct the image after adding noise back to the image before adding noise through learning, and its loss function can be expressed as: for the observed data x, it is hoped to find the mean and variance to maximize the likelihood function value, as shown in Formula (1).

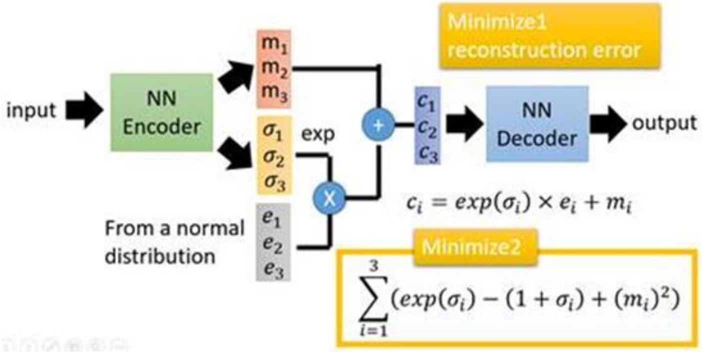
Minimize1
reconstruction error
input
NN Encoder
m_1 m_2 m_3
σ_1 σ_2 σ_3
exp
+
X
c_1 c_2 c_3
NN Decoder
output
From a normal distribution
e_1 e_2 e_3
$c_i = exp(\sigma_i) \times e_i + m_i$
Minimize2
$\sum_{i=1}^{3}(exp(\sigma_i) - (1+\sigma_i) + (m_i)^2)$

Figure 1: Automatic variational encoder structure.

$$L_{VAE}(E,G) = \lambda_1 KL(q(z\,|\,x)\,||P(z)) - \lambda_2 E_{q(z\,|\,x)}[log\,P\,(x|z)] \tag{1}$$

Where KL is a divergence function; The two hyperparameters λ_1 and λ_2 are the weights of two losses respectively; z is the decoder input; x is the encoder input; E is the expected value of the distribution q (z | x).

The generated countermeasure network consists of a generator G and a discriminator D, wherein the structure of the generator is the structure of the decoder in VAE, and the generator generates two types of images, namely, a reconstructed image and a converted image, and the countermeasure training is to train the images in the converted stream. The loss function is shown in Formula (2).

$$L_{GAN} = \lambda_0 E_{x\sim p(x)}[logD(x)] + \lambda_0 E_{z\sim q(z\,|\,x)}[log(1 - D(G(z)))] \tag{2}$$

Where D is the discriminator; G is the generator.

The cyclic consistent loss in the unsupervised image transformation model is to ensure that the image can be transformed back after cross-domain transformation, and the loss function is shown in Formula (3).

$$L_{cc}(E_1, G_1, E_2, G_2) = \lambda_3 KL(q_1(z_1\,|\,x_1)\,||P(z)) + \lambda_3 KL(q_2(z_2\,|\,x_2)\,||P(z)) - \lambda_4 E_{q(z\,|\,x)}[logP_{G1}(x_1|z_2)] \tag{3}$$

Therefore, the overall loss function of the image transformation model is composed of three parts: encoding-decoding loss, generation confrontation loss, cyclic consistency loss.

Taking 20,000 images from GTA5 data set as the source domain training set and 2,975 images from Cityscape data set as the target domain training set, after training for 150K cycles, 24,966 images in the source domain are transformed into the target domain by using the trained model. The conversion process from the source domain to the target domain is shown in Figure (2). The first line is the original images of four different samples, the second line is the reasoning result after training the model for 50,000 times, and the third line is the reasoning result after training the model for 100,000 times. As can be seen from the figure, with the increase of training times, the generated image is closer to the real scene.

The following tests the trained transformation model, and the test results are as follows.

As shown in Figure (3). In the figure, the first line is a partial sample from the GTA5 data set, and the second line is the sample after it is converted into the target domain.

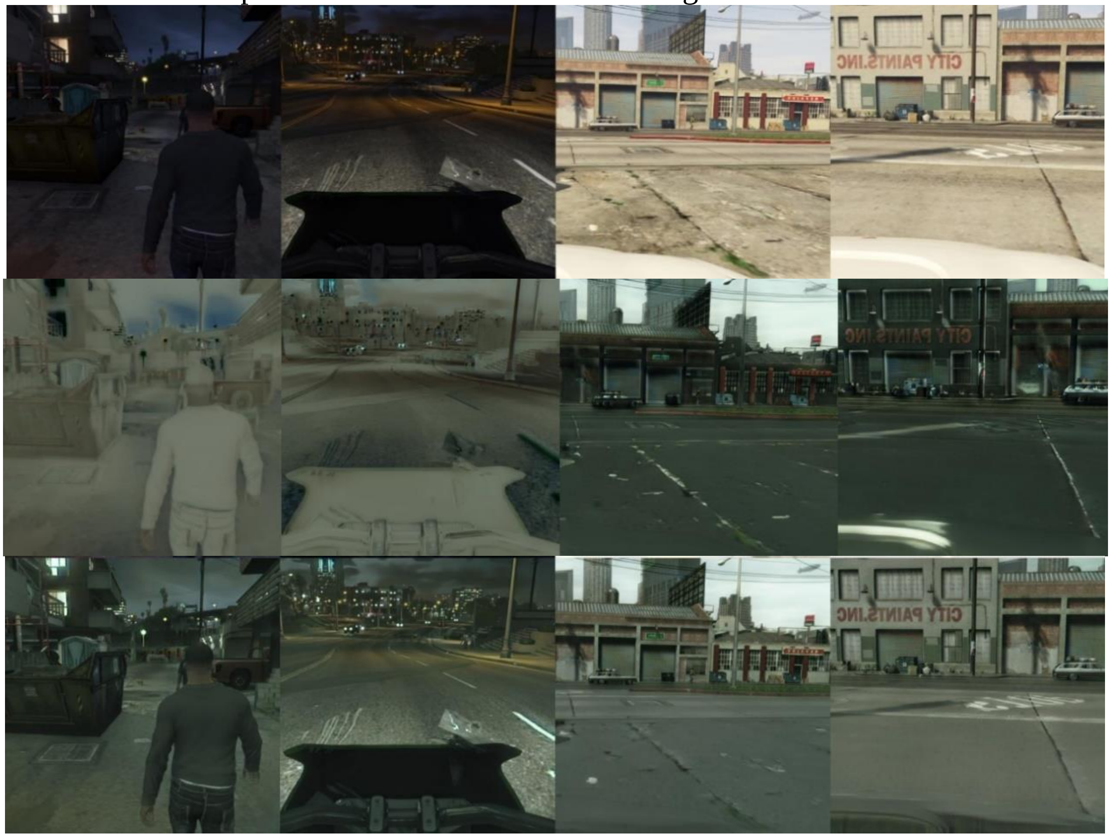

Figure 2: Effect of forward conversion in image transformation network training process.

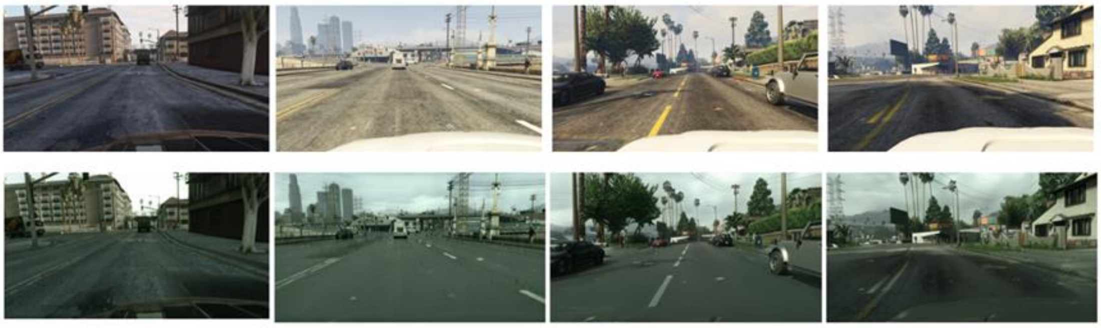

Figure 3: Forward conversion test results.

As can be seen from the Figure (3), the transformed picture is clearer and more realistic, and the information in the picture is more complete. Therefore, the transformed GTA5 data set is named GTA2City, and it is used as the intermediate domain for the second stage of domain adaptation.

3.2. Unsupervised Image Transformation Model

After completing the first phase, namely the image transformation phase, the image transformation framework and the single-phase domain adaptive method are integrated. In the first phase, the network input is the source domain GTA5 sample and the target domain Cityscape sample, and the output is the intermediate domain GTA2City sample. In the second stage, the network input is intermediate samples, segmented labeled and unlabeled target domain training samples, and the network output is the classification result of pixel points and the judgment result of discriminator.

500 test samples from Cityscape verification set are used to test the model after confrontation training, and some test results are shown in Figure (4).

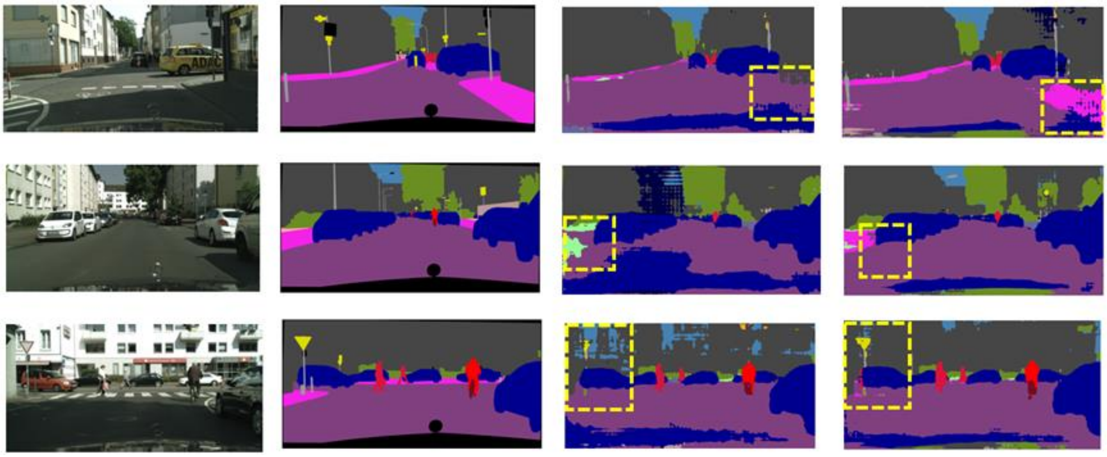

Figure 4: Cross-domain segmentation results of two-step domain adaptation.

As shown in the figure(4), the first column is the original image of three samples, the second column is the truth label corresponding to the first column, and the third and fourth columns are the segmentation results of the one-step domain adaptive method and the two-step domain adaptive method respectively. We can see from the area enclosed by the dotted line in the figure that the segmentation effect of the two-step domain adaptive method is better than that of the one-step domain adaptive method.

Compared with the single-step domain adaptive method, the average intersection ratio of the two-step domain adaptive method is increased by about 5.5 percentage points, and the advantages of the two-step domain adaptive method are more obvious in the categories of pavement, sidewalk and pedestrian.

3.3. The Training Mechanism of Sample Weight Based on Image Quality Evaluation.

As shown in Figure (3), the generated images of these samples are of high quality, so we can regard them as a reliable intermediate domain. But we can't guarantee that the quality of all generated images is very high. As shown in Figure (5), some fake leaves appear out of thin air in the generated images. The goal of image quality assessment is to design a model to quantify the image quality with as little prior knowledge as possible. Therefore, this section uses the evaluation method based on probability model to score the generated intermediate domain images, so as to reduce the influence of low-quality generated images on the final segmentation results.

The image quality is evaluated by calculating the distance between the image parameters to be evaluated and the pre-established standard model parameters. The distance formula is calculated as shown in Formula (4).

$$D(v_1, v_2, \Sigma_1, \Sigma_2) = \sqrt{((v_1 - v_2)^T (\frac{\Sigma_1 + \Sigma_2}{2})^{-1} (v_1 - v_2))} \quad (4)$$

In the formula, v_1 and v_2 respectively represent the average value of the normal image and the image to be measured, and Σ_1 and Σ_2 respectively represent their covariance matrices.

According to Formula (4), the worse the quality of the generated image, the greater the D, that is, the greater the distance between the generated image and the natural image. In this paper, we need to give this kind of low-quality generated image a smaller weight to make it have a lower impact on the two-step domain adaptive training. On the contrary, this paper needs to give more weight to the generated images with high quality.

Therefore, firstly, this section calculates the NIQE distance for each sample, and then normalizes it to make the picture quality evaluation score in the range of 0 to 1, and uses the evaluation score as the sample weight of the two-stage training to guide the two-stage training process. The normalization treatment is shown in Formula (5).

$$w(I) = 1 - \frac{D - D_{min}}{D_{max} - D_{min}} \tag{5}$$

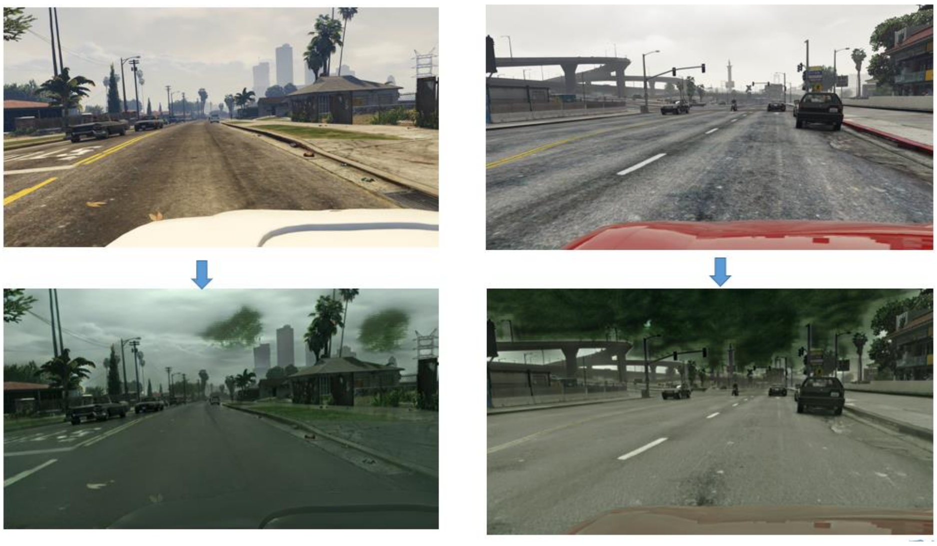

Figure 5: Sample examples of distortion after image transformation.

Next, the GTA5 sample and its labeled and unlabeled Cityscape training set samples are input into the model for training, and then the Cityscape verification set (500 samples) is segmented by using the trained model. The segmentation results of three samples are shown in Figure (6). The first column to the fifth column are the original image of the sample, the truth label, the domain-free adaptive segmentation result, the single-step domain adaptive segmentation result, the two-step domain adaptive segmentation result and the two-step domain adaptive segmentation result with sample weight.

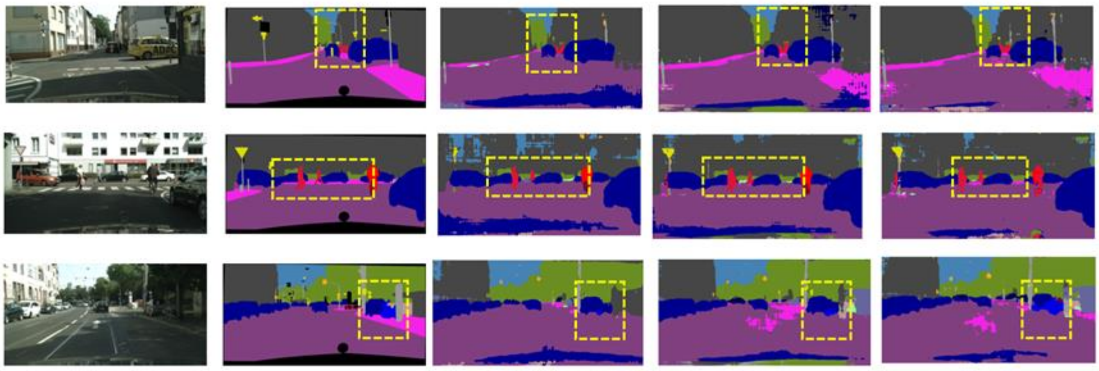

Figure 6: Cross-domain segmentation results by introducing sample weight.

We can see in the framed areas of these three samples that only the two-step adaptive model with sample weight can completely segment the street lamp of the first sample, the sidewalk of the second sample and the motorcycle of the third sample.

The average intersection ratio of the two-step domain adaptive segmentation method with sample weight is 44.62%, which is about 6 percentage points higher than that of the single-step domain adaptive segmentation method.

4. Experiment and Result Analysis

The data acquisition vehicle is SAIC Roewe Ei5, and the video acquisition equipment is Mijia driving recorder. The data acquisition prototype vehicle and video acquisition equipment are shown in Figure (7).

Figure 7: Data acquisition sample car and video acquisition equipment.

In this section, a captured daytime driving video is made into a data set, which is used to test the adaptive domain segmentation algorithm. As shown in Figure (8), lines 1 to 5 are the segmentation test results of the first frame, the 100th frame, the 300th frame, the 400th frame and the 600th frame in the video, and from left to right are the original image of the sample, the segmentation result of domain-free adaptation, the segmentation result of single-step domain adaptation and the segmentation result of two-step domain adaptation.

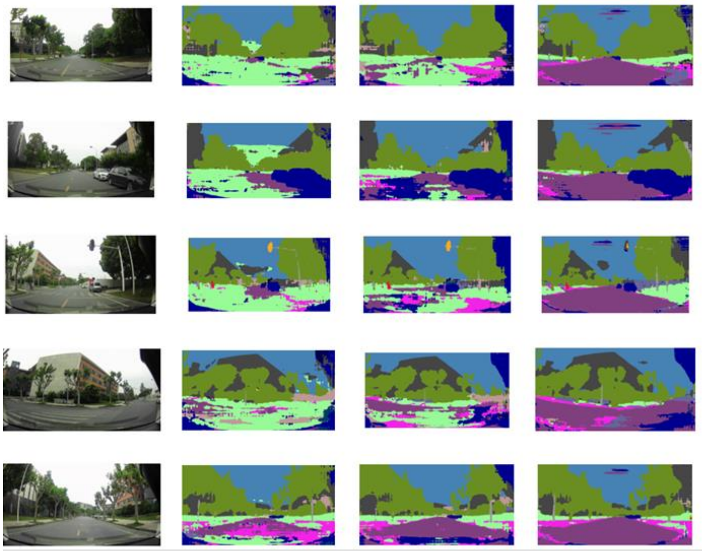

Figure 8: Real scene video test results.

From the test results, it can be seen that compared with the single-step domain adaptive method, the two-step domain adaptive method has better segmentation effect on roads, sidewalks and grasslands.

Taking Cityscape data set as the source domain and the night scene data set collected by the real vehicle as the target domain, the segmentation result is shown in Figure (9), and the first to fourth lines are the original image of the night driving scene, the segmentation result of the domain-free adaptive method, the segmentation result of the two-step domain adaptive method and the segmentation result of the improved two-step domain adaptive method respectively.

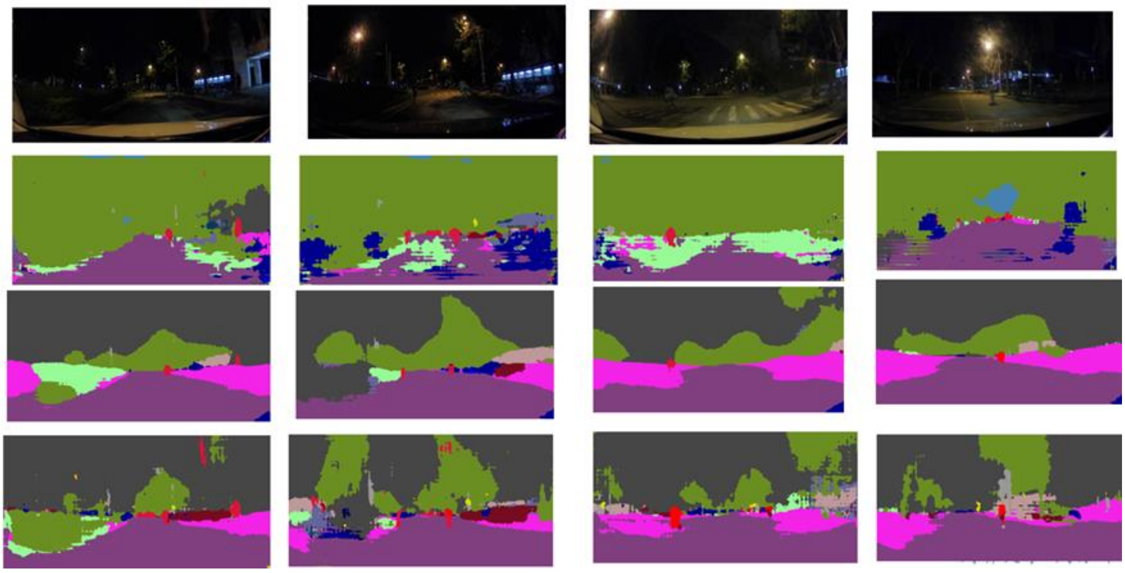

Figure 9: Semantic segmentation result of night driving scene.

As can be seen from the Figure (9), the segmentation effect of the domain-free adaptive method is poor, and the road segmentation by the two-step domain adaptive method is clear. The improved two-step domain adaptive method can even segment street lamps and bicycles.

5. Conclusion

In this paper, a two-stage domain adaptation method based on image transformation is proposed, which decomposes the adaptation problem from source domain to target domain into two stages: in the first stage, the source domain image is transformed to target domain, and the obtained result is used as the intermediate domain; In the second stage, a single-stage domain adaptive method based on antagonistic learning is used at the feature level (output space). In this paper, the same backbone network and segmentation model are used, and the mIOU is improved by about 5.5% compared with the results of single-step domain adaptive method in the same sample test.

Based on the two-stage domain adaptive method, this paper introduces the scoring mechanism of image quality evaluation, adjusts the weight of the second-stage training samples, and improves the segmentation mIOU by about 6% compared with the single-stage domain adaptive method.

Finally, this paper uses real vehicles to collect driving video data, including daytime and nighttime driving scenes, and verifies the semantic segmentation effect of this method by adapting the synthetic data field to the real driving scene field with the daytime data of campus scenes. Then Cityscape is used as the source domain and campus night driving scene as the target domain, which verifies the adaptive segmentation effect of this method from daytime data domain to night data domain, thus verifying the unsupervised adaptive semantic segmentation method in different applications.

References

[1] Klare B F, Klein B, Taborsky E, et al. Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A[C]// IEEE Computer Vision and Pattern Recognition (CVPR), 2015.

[2] Deng W, Liang Z, Kang G, et al. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification [J]. 2017.

[3] Wei L, Zhang S, Wen G, et al. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[4] Zhong Z, Liang Z, Zheng Z, et al. Camera Style Adaptation for Person Re-identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[5] Chen Y, Li W, Sakaridis C, et al. Domain Adaptive Faster R-CNN for Object Detection in the Wild[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[6] Inoue N, Furuta R, Yamasaki T, et al. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[7] Kim T, Jeong M, Kim S, et al. Diversify and Match: A Domain Adaptive Representation Learning Paradigm for Object Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[8] Roychowdhury A, Chakrabarty P, Singh A, et al. Automatic Adaptation of Object Detectors to New Domains Using Self-Training[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[9] Vincent O. Rebecca, and Olusegun Folorunso. "A descriptive algorithm for sobel image edge detection." Proceedings of informing science & IT education conference (InSITE). Vol. 40. 2009.

[10] Wang X, Cai Z, Gao D, et al. Towards Universal Object Detection by Domain Attention [J]. IEEE, 2019.

[11] Zhu X, Pang J, Yang C, et al. Adapting Object Detectors via Selective Cross-Domain Alignment[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[12] Wang T, Zhang X, Yuan L, et al. Few-shot Adaptive Faster R-CNN[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[13] Pan F, Shin I, Rameau F, et al. Unsupervised Intra-Domain Adaptation for Semantic Segmentation through Self-Supervision [J]. Conference on Computer Vision and Pattern Recognition (CVPR), 2020.