An Adversarial Self-Learning Method for Cross-City Adaptation in Semantic Segmentation

Huachen Yu and Jianming Yang

***Abstract*—Semantic segmentation is an important task in the visual system of self-driving cars. The semantic segmentation models based on the CNN (Convolutional Neural Network) trained with the large numbers of annotated labels may not work well at the environments different from the training sets due to the domain gap between the train and test domains. Just for the reduction of the distance between the source and target domains, domain adaptation methods are proposed for the unsupervised training with the unlabeled target domain. Not only the reduction of the domain-shift, but we also propose the self-learning method to enhance the predicted probabilities of the target domain. To gain more accurate probability maps of the target domain generated from the segmentation model which is trained by the source domain, we propose the adversarial self-learning method which is consists of the domain adaptation part and self-learning part. The adversarial self-learning method can maximize the predicted probabilities for the probability maps of the target domain gained from the segmentation model which is adapted with the domain adaptation method before the self-learning. With the Cityscapes to NTHU cross-city adaptation experiments, we can see that the adversarial self-learning method can achieve state-of-the-art results compared with the domain adaptation methods proposed in the recent researches.**

***Index Terms*—Semantic segmentation, domain adaptation, adversarial self-learning, cross-city adaptation.**

I. Introduction

With the visual system for self-driving cars, we can realize the line and road detection [1], traffic sign recognition [2], depth estimation [3], objection detection [4] and semantic segmentation [5] based on the image processing techniques. Just for the understanding of the urban scenes, semantic segmentation plays a significant role in the visual system. Different from the image recognition which is an image-wise classification problem, semantic segmentation is a pixel-wise classification task which gives each pixel of the image a label. With the improvement of the CNN architectures, the performance of the semantic segmentation system [6]-[15] was significantly increased in a few years. For the supervised leaning for the semantic segmentation system based on the CNN architectures, a large number of high quality annotated images [16] are needed. For the training of the semantic segmentation system used for the urban scene understanding, Cityscapes [16] datasets and Mapillary Vistas [17] datasets contained thousands of high-annotated images from multi-cities from the real world are provided. As the high cost for the annotation of the pixel-level labels, the synthetic datasets [18], [19] which use the labels rapidly generated from the computer games for the semantic segmentation models training are provided. When we use the semantic segmentation model pre-trained with the real-world datasets or synthetic datasets to predict the images from the scenes which are not in the training datasets, the semantic segmentation models may not give good performance due to the domain shift [20]-[23] between the training datasets and the testing datasets. Retraining the models with the testing datasets is impossible with the not enough annotated labels. To deal with the domain shift problems for the semantic segmentation system, domain adaptation methods [24]-[29] based on the GANs [30], [31] (Generative Adversarial Networks) which are used to reduce the divergence between the two domains (training datasets and testing datasets) are proposed. As a proposed method, we choose the probability enhancement for the object prediction results which are the outputs gained from the semantic segmentation system (probability maps) of the target domain (testing dataset) to adapt the segmentation models trained with the source domain (training datasets). Our contributions in this paper can be introduced as follows:

As a self-learning method, we calculate the cross-entropy loss with the pseudo labels gained from the probability maps of the target domain to enhance the object prediction probabilities.

To gain accurate pseudo labels for the self-learning, we use the outputs space domain adaptation method used the GANs to reduce the domain gap for the probability maps of the target domain.

We propose an adversarial self-learning method which is a combination of the domain adaptation method and the self-leaning method. We implement the proposed method for the real world cross-city adaptation. The experiments show that the proposed method can achieve state-of-the-art results.

II. Related Works

In this section, we introduce some semantic segmentation systems based on CNN models and some domain adaptation methods based on the GANs in recent researches.

Semantic Segmentation. The semantic segmentation systems achieve rapid development with the CNN models in recent years. Like the architecture of the FCN [6] (fully convolutional network), the segmentation system contains two parts, the feature extractor, and the classifier module. The feature extractors use the image recognition models like the AlexNet [32], VGGNet [33], GoogleNet [34], and ResNet [35], etc. pre-trained with the ImageNet [36] and

Manuscript received December 5, 2019; revised May 1, 2020.

Huachen Yu and Jianming Yang are with the Department of Mechanical Engineering, Meijo University, Nagoya, Japan (e-mail: huachen_yu@yahoo.com, yang@meijo-u.ac.jp).

doi: 10.18178/ijmlc.2020.10.5.986

Microsoft COCO [37] datasets to extract the feature maps for the images from the segmentation datasets. The classifier modules use the deconvolution layers for the pixel-wise classification with the consistency of the channels and sizes of the probability maps based on the extracted feature maps. Unlike the FCN [6] model, the U-Net [38], SegNet [15] models proposed the architecture consist of the encoder and decoder modules for the segmentation system. Instead of the pre-trained feature extractors, the encoder-decoder models which are used to generate probability maps from the input images directly are trained with the semantic segmentation datasets like end-to-end systems. Just like the architecture of FCN, the DeeplabV2 [7] uses the pre-trained ResNet101 [35] as the backbone for the feature extraction. Instead of the deconvolution layers, DeeplabV2 use the ASPP [7] (Atrous Spatial Pyramid Pooling) as the classifier module which uses the dilated convolution and the multi-filters with different rates to gain image spatial context with multi-scales.

Domain Adaptation. The domain adaptation methods based on GANs [30] used for the semantic segmentation systems in recent researches can be divided into three classes: the feature adaptation, the outputs adaptation, and the image adaptation. The feature adaptation methods [27], [28] use the discriminator to calculate the distance of the distributions of the feature maps which are extracted from the images of source and target domains with the pre-trained backbones as VGG16 [33], ResNet101 [35]. With the reduction of the distance of the feature maps of the source and target domains, the classification results of the target domain based on the feature maps can be similar to the source domain. As systematic outputs can be generated with the semantic segmentation system, the outputs adaptation methods [24], [29] directly use the discriminator to calculate the divergence of the probability maps which are the segmentation system outputs. With the adversarial training for the segmentation networks, the distributions of the probability maps from the target and source domains can be as close as possible. As the difference of image styles is the reason for the domain shift of the semantic segmentation system, the image adaptation methods [25], [26] use the GANs for image-to-image translation. As the images generated from the target domain images based on the style transfers which use the GANs can gain similar styles with the source target images, the domain gap for the segmentation system can be reduced.

III. Proposed Method

In this paper, we proposed a new algorithm consists of adversarial learning and self-learning methods for the output space gained from the segmentation network to deal with the domain shift problem between the source domain and target domain for the semantic segmentation system. In this section, we explain the algorithm flow for the proposed method and the loss function for the system optimization in details.

A. Architecture Overview

As shown in Fig. 1, the proposed method can be divided into two parts: the segmentation network G and the discriminator module D. Just for a semantic segmentation system, input images and annotated labels as the ground truth from source domain are used to calculate the cross-entropy loss to train the weights of the segmentation network G as a supervised learning. For the domain adaptation part which is used to reduce the domain-shift between source and target domains, the adversarial losses for target domain images as the JS (Jensen–Shannon) divergence calculated by the discriminator module proposed by Ian Goodfellow [30] can be used to fine-turn the trained weights of the segmentation network. As the divergence between the outputs space of the source and target domains been reduced by the adversarial loss, the probability maps of the target domain as the outputs of the segmentation network can be used for a self-learning process. To enhance the confidence of the objects probabilities from the probability maps of the target domain, the cross-entropy loss between the probability maps and the pseudo labels gained from the probability maps can be calculated for the self-learning method used to adapt the weights of the segmentation network. As the generative adversarial learning, the adversarial loss from the discriminator D can be used to adapt the segmentation network G, the weights of the discriminator D should be trained with the probability maps of the source and target domains to distinguish the domains of the outputs space generated from the segmentation network G.

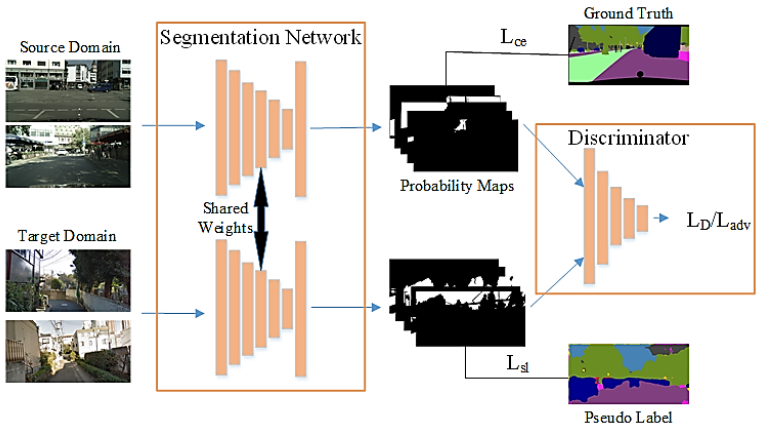

Fig. 1. Overview of the proposed algorithm. The proposed method can be composed of segmentation network and the discriminator module. We use the given images and annotated labels from the source domain as a supervised training for the segmentation network. To reduce the domain-shift between the source and target domains, images from the target domain can be used to adapt the segmentation network as an unsupervised training. We use the probability maps of source and target domains gained from the segmentation network to train the discriminator module.

B. Loss Function

Segmentation Network Training. As introduced in section A, we use the images I_s and the annotation labels Y_s from the source domain to train the segmentation network G as a supervised learning. We use the images I_t from the target domain to adapt the segmentation network G to reduce the domain-shift between the source and target domains. As $I_s, I_t \in R^{H\times W\times 3}$ (H and W are the height and width of the images), the overview loss function can be expressed as:

$$\mathrm{L}(I_s, I_t) = L_{ce}(I_s) + \lambda_{adv} L_{adv}(I_t) + \lambda_{sl} L_{sl}(I_t) \quad (1)$$

where $L_{ce}(I_s)$ is the cross-entropy loss between the source domain images I_s and the annotated labels Y_s, $L_{adv}(I_t)$ is the adversarial loss for the probability maps of the target domain images I_t calculated by the discriminator D, and $L_{sl}(I_t)$ is the self-learning loss which is the cross-entropy loss calculated between the probability maps gained from target domain images I_t and the pseudo labels. The λ_{adv} and λ_{sl} are the weights for the adversarial loss $L_{adv}(I_t)$ and the self-learning loss $L_{sl}(I_t)$.

Cross-entropy loss for a supervised learning. We use the images I_s and annotated labels Y_s to train the weights of the segmentation network G. As C is the number of categories, $\mathrm{G}(I_s) \in R^{H\times W\times C}$ is the outputs of the segmentation network as probability maps. Before the calculation of the loss function, the probability maps $\mathrm{G}(I_s)$ should be normalized with a softmax layer. With the definition of the normalization probability maps $\widehat{\mathrm{G}}(I_s)$ as $\widehat{\mathrm{G}}(I_s) = \mathrm{Softmax}(\mathrm{G}(I_s))$, we can define the cross-entropy loss $L_{ce}(I_s)$ which is based on the source domain as:

$$L_{ce}(I_s) = -\sum_{h,w}\sum_{c\in C} Y_s^{(h,w,c)} \log(\hat{G}(I_s)^{(h,w,c)}) \quad (2)$$

where the $Y_s^{(h,w,c)}$ is the one-hot encoder for the annotated labels Y_s.

The adversarial loss for the unsupervised learning. We use the Discriminator D which can distinguish the domains of the probability maps which are the outputs of the segmentation network G to calculate the adversarial loss L_{adv}. As the inputs for the discriminator D, we should use the normalization probability maps $\widehat{\mathrm{G}}(I_t)$ instead of the probability maps $\mathrm{G}(I_t)$ generated from the target domain images I_t. As the discriminator D is a classifier to distinguish the probability maps domains, we label the probability maps of the source domain with 1, the adversarial losses for I_t can be defined as:

$$L_{adv}(I_t) = -\sum_{h,w} \log(D(\hat{G}(I_t))^{(h,w)}) \quad (3)$$

As we minimize the adversarial loss L_{adv} to adapt the segmentation network G, the distribution of the target domain probability maps can be close to the source domain.

Self-learning loss for an unsupervised learning. As a self-learning method, we use the cross-entropy loss between the probability maps $\mathrm{G}(I_t)$ gained from the target domain images I_t and the pseudo labels $\hat{Y}_t$ gained from the probability maps $\mathrm{G}(I_t)$ to maximize the probabilities of objects (road, sidewalk, tree, etc.) of the target domain images pixels. Dealing with the consistency of the adversarial loss, we use the normalization probability maps $\widehat{\mathrm{G}}(I_t)$ instead of the probability maps $\mathrm{G}(I_t)$ for the calculation of the self-learning loss. The pseudo labels of target domain $\hat{Y}_t$ can be gained from the normalization probability maps $\widehat{\mathrm{G}}(I_t)$ with an argmax function, which can be defined as $\hat{Y}_t = argmax(\widehat{\mathrm{G}}(I_t))$. The self-leaning loss calculated with the cross-entropy loss between $\widehat{\mathrm{G}}(I_t)$ and $\hat{Y}_t$ can be defined as:

$$L_{sl}(I_t) = -\sum_{h,w}\sum_{c\in C} \hat{Y}_t^{(h,w,c)} \log(\hat{G}(I_t)^{(h,w,c)}) \quad (4)$$

Discriminator Network Training. The discriminator D is a two-class classifier to distinguish the domains of the normalization probability maps $\widehat{\mathrm{G}}(I_s)$ and $\widehat{\mathrm{G}}(I_t)$. The overview loss used to train the weights of discriminator D with $\widehat{\mathrm{G}}(I_s)$ and $\widehat{\mathrm{G}}(I_t)$ can be expressed as:

$$L_D(I_s, I_t) = -\sum_{h,w} \log\left(D\left(\hat{G}(I_s)\right)^{(h,w)}\right) + \log\left(1 - D\left(\hat{G}(I_t)\right)^{(h,w)}\right) \quad (5)$$

As the generative adversarial learning, the minimization of the discriminator loss $L_D(I_s, I_t)$ can gain a JS divergence [30] between the distributions of the probability maps from source and target domains.

IV. Experiments and Results

In this section, we use the results of the experiments to validate the effectiveness of domain adaptation method for the semantic segmentation system proposed in this paper. For the real world cross-city adaptation experiments, we introduce the system network architecture (segmentation network G and discriminator D), the setting of the parameters and optimizer functions for the domain adaptation system, environments and datasets of the experiments, discussion on the results in details.

A. Network Architecture

Segmentation Network G. We use the Deeplabv2 [7] as the base architecture for the segmentation network G in the experiments. For the Deeplabv2 model, we adopt the Resnet101 [35] network which has been pre-trained with the ImageNet dataset as the backbone of segmentation network G for the feature extraction. We use the ASPP module as the decoder for G to gain the probability maps from the input feature maps. As of last, for the consistency of the input size, we use an up-sampling layer to resize the probability maps from the ASPP [7] module.

Discriminator D. To pay attention to the local patches from the input probability maps, we use the PatchGAN [39] as the base architecture for the discriminator D. The discriminator D is composed of 5 convolution blocks with the output channels as {64,128,256,512,1}. Each block used in PatchGAN consists of a convolution layer with the kernel size set to 4, stride size set to 2, padding set to 1, and a LeakyReLU [40] layer used as the activation layer with the negative slope set to 0.2.

B. Parameters and Optimizers

For the weight of the adversarial loss λ_{adv}, as indicated in the AdaptSegNet [24], we adopt 0.001 to gain a sensitive effect for the adaptation. We also choose 0.001 for the weight λ_{sl} of the self-learning part to keep the consistency of adversarial learning. We choose the SGD [41] (Stochastic

Gradient Descent) optimizer for the segmentation network G with the parameters' initial learning rate set to 2.5×10^{-4}, momentum set to 0.9, and weight decay set to 10^{-4}. For the discriminator D, we use Adam optimizer [41] with the initial learning rate set to 10^{-4}. Just for training, the learning rates of the optimizers used for the G and D are decreased with the polynomial decay as the power set to 0.9.

C. Datasets and Environments

Datasets. In the experiments, we use the cityscapes [16] dataset as the source domain and cross-city dataset [28] as the target domain to implement domain adaptation for the semantic segmentation system. The cityscapes [16] dataset contains 5000 high quality pixel-wise annotated images from 50 cities around Europe. The dataset is focused on the urban street scenes and labeled with 30 classes. We only use the training set contained 2975 images from the cityscapes dataset consists of training, testing and validation parts for the segmentation network training. The cross-city NTHU [28] dataset is used to show the different appearance from the cityscapes dataset collected from four cities Rome, Rio, Tokyo, and Taipei. For each city, 3200 unannotated images are used to adapt the domain shift and 100 annotated images used to validate the adaptation effect of the system. In this paper, to prevent the domain adaptation system from the over-fitting problems, we choose the cities Rio, Tokyo, and Taipei which are not the European cities for the adaptation experiments. As the image size for the experiments, the height is set to 256, the width is set to 512.

Experiments environment. Our proposed adversarial self-learning method is implemented with the Pytorch framework. We train the segmentation network G and discriminator D with NVIDIA GTX 1080ti GPU for 100000 iterations took about 12 hours. We use the testing set from the cross-city dataset to validate the system and save the weights every 3000 iterations.

D. Overview Results

In this paper, as the cross-city dataset is labeled with 13 classes, we calculate the mIoU [42] (Mean Intersection over Union) which is the mean IoU of the 13 classes as the metric for the semantic segmentation system. Table I presents the results for the three cities' (Rio, Tokyo, and Taipei) segmentation performance transferred from the cityscapes dataset. In Table I, the SW, BLDG, TL, TS, VEG, Motor. are used to stand for Sidewalk, Building, Traffic Light, Traffic Sign, Vegetation, and Motorbike; the AL and SL are used to stand for the adaptation learning and self-learning. With the results of the experimentations, our proposed adversarial self-learning method (AL+SL) in this paper can be compared with the feature adaptation method (AL(Feature)) mentioned in the [28] and the output space adaptation method (AL(Outputs)) proposed by [24] to show the advantages when dealing with the domain shift problems for the segmentation system. As mentioned in the previous researches, the deep network can achieve better feature representation and segmentation results, we used the ResNet101 as the backbone for all the experiments. With Table I, we can see that both the adaptation of the feature map and the output space can gain effective performance. With no adaptation operation, the domain adaptation can reduce the domain shift in the segmentation system. To compare with the feature adaptation method to reduce the divergence between the feature maps of the source and target domains, directly reducing the divergence of the pixel-wise classification results used the outputs space adaptation method achieves the better mIoU results. The proposed self-learning method (SL) which is used the cross-entropy loss with the pseudo labels can reduce the domain shift for the segmentation system based on the results of the experiments. For real-world cross-city adaptation, we can see that the self-learning method gains better performance compared with the domain adaptation methods from tabel1. As the pseudo labels used in the self-learning method are gained from the output probability maps, we proposed the adversarial self-learning method which uses the outputs space adaptation to reduce the divergence between the probability maps of the source and target domains before the self-learning for the target outputs. From tabel1, we can see that the adversarial self-leaning method proposed in this paper achieves **state-of-the-art** results compared with the baseline (with no adaptation) and the domain adaptation methods proposed in recent years.

TABLE I: THE RESULTS OF THE CITYSCAPES TO CROSS-CITY ADAPTATION FOR SEMANTIC SEGMENTATION SYSTEM

		Cityscapes - Cross-City													
City	Method	Road	SW	BLDG	TL	TS	VEG	Sky	Person	Rider	Car	Bus	Motor.	Bicycle	MIoU
Tokyo	Baseline	67.79	18.35	56.80	**1.12**	4.20	68.20	39.43	11.22	**2.95**	56.04	0.33	1.72	26.55	27.28
	AL(Feature)[28]	**73.12**	**22.90**	61.80	0.37	1.75	66.76	60.05	11.93	0.57	56.45	0.00	0.34	19.58	28.84
	AL(Outputs)[24]	68.53	16.04	64.84	0.58	4.47	**69.05**	64.80	16.56	0.64	53.74	3.61	**4.46**	20.90	29.86
	SL	71.34	20.11	62.39	0.97	4.72	68.65	65.21	15.21	0.26	**58.63**	0.28	3.91	22.95	30.36
	AL+SL	69.69	17.97	**66.73**	0.85	**4.72**	68.97	**71.51**	**20.44**	0.99	55.65	**3.69**	3.71	**29.97**	**31.91**
Taipei	Baseline	49.03	14.35	63.26	**1.93**	2.86	58.15	58.57	5.81	0.66	34.10	3.94	**14.06**	2.52	23.79
	AL(Feature)[28]	44.83	12.52	66.81	1.71	2.87	58.06	52.65	**6.64**	0.40	36.13	1.99	13.21	4.50	23.26
	AL(Outputs)[24]	52.16	**15.96**	70.36	1.40	**3.36**	61.24	69.09	5.44	1.59	30.78	**14.70**	1.56	2.12	25.37
	SL	46.97	15.66	67.22	1.80	2.83	**61.62**	67.25	6.62	**3.35**	**39.91**	6.18	9.19	**6.30**	25.76
	AL+SL	**56.01**	15.69	**70.98**	1.64	2.89	59.24	**71.13**	6.55	1.47	33.74	12.98	9.83	1.82	**26.46**
Rio	Baseline	38.29	16.76	58.63	0.53	2.05	67.83	53.58	15.46	1.62	44.36	6.26	7.63	7.30	24.64
	AL(Feature)[28]	37.08	19.12	59.15	**0.64**	2.00	68.51	51.16	16.23	**3.81**	42.47	4.96	5.34	7.32	24.45
	AL(Outputs)[24]	**56.34**	18.36	61.31	0.31	**2.39**	66.83	52.88	18.31	0.75	42.26	**8.68**	1.28	8.08	25.98
	SL	48.92	**19.18**	61.68	0.48	2.18	**69.58**	59.84	**19.40**	1.12	**46.45**	7.01	3.31	10.57	26.90
	AL+SL	50.31	17.78	**66.62**	0.36	1.85	67.65	**64.56**	19.09	0.57	43.90	8.25	**11.19**	**11.04**	**27.93**

In Fig. 2, we select some semantic segmentation outputs from the three cities in the experiments to show the effective

performances of the adversarial self-learning method proposed in this paper. In Fig. 2, the GT is used to strand for the ground truth (the human-annotated data), the results of our proposed method are compared with the results from the baseline and the outputs space adaptation methods. To show the effectiveness of the adversarial self-learning method, we pay attention to the regions with red bounding boxes. From the results of Tokyo, we can see that the region of the sky can't be well classified only with the baseline, and the domain adaptation methods for the segmentation system can resolve the domain shift problems. With the results of Taipei, from the middle region of the outputs, we can see that combined with the self-learning method which can enhance the probability maps, the existed persons can be detected compared with the outputs adaptation method. With the middle-left regions of the results from Rio, the adversarial self-leaning method reduces the noise region for the predicted trees compared with the baseline and outputs adaptation method. With all the experiment results, the adversarial self-learning method can achieve **state-of-the-art** results.

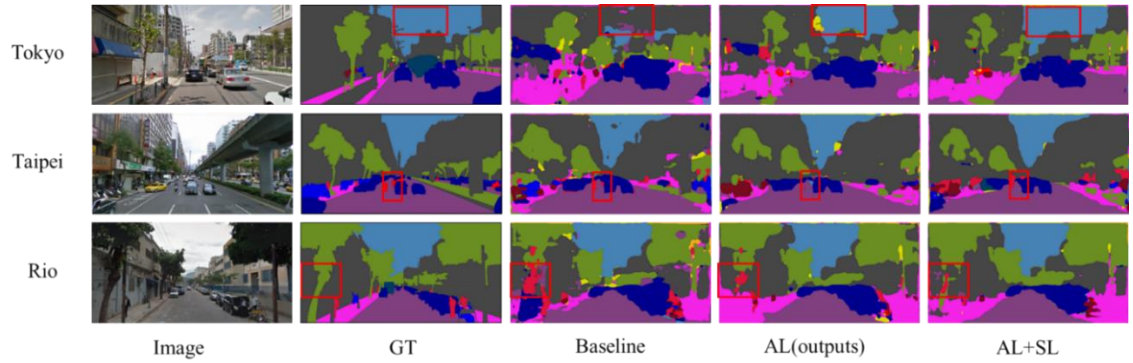

Fig. 2. The example results generated from the cityscapes to cross-city adaptation system for the three cities. Each city contains the original image, the ground truth and the predicted label maps generated from the compared adaptation methods and our proposed method.

V. Conclusion

In this paper, to deal with the domain shift problems of the different cities for the semantic segmentation system, we proposed an unsupervised learning method only with the annotated source images. The proposed method used a self-leaning method with the pseudo labels to enhance the confidence of outputs probability maps and combined with the outputs domain adaptation to enhance the confidence of the pseudo labels. With the cityscapes to cross-city experiments, our method can achieve state-of-the-art results for the domain shift problems. We hope that our proposed method can gain better performance with the synthetic to real segmentation tasks.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Huachen Yu designed the study, performed the experiments, analyzed the data and wrote the paper, Jianming Yang supervised the research and revised the paper; all authors had approved the final version.

References

[1] W. Farag and Z. Saleh, "Road lane-lines detection in real-time for advanced driving assistance systems," *in Proc. International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 2018.

[2] R. Q. Qian, Y. Yue, F. Coenen, and B. L. Zhang, "Traffic sign recognition with convolutional neural network based on max pooling positions," in *Proc. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2016, pp. 578-582.

[3] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 3828-3838.

[4] X. L. Wang, A. Shrivastava, and A. Gupta, "A-Fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2606-2615.

[5] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4151-4160.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFS," in *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 1520-1528.

[9] L.-C. Chen, Y. Yang *et al.*, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640-3649.

[10] Y. Li, H. Z. Qi, J. F. Dai, X. Y. Ji, and Y. C. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp 2359-2367.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. European Conference on Computer Vision*, 2018.

[12] 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

[13] C. X. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. Yuille, F.-F. Li, "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 82-92.

[14] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, Hui Huang, "Multi-Scale Context Intertwining for Semantic Segmentation," *in Proc. European Conference on Computer Vision*, 2018.

[15] 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.

[16] M. Cordts, M. Omran, S. Ramos, and T. Rehfeld, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213-3223.

[17] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 4990-4999.
[18] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. European Conference on Computer Vision*, 2016.
[19] G. Ros, L. Sellart, J. Materzynska, D.Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4321-4330.
[20] Y. Zhang, P. David, and B. Q. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 2020-2030.
[21] Y. Zou, Z. D. Yu, B. V. K. V. Kumar, and J. S. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. European Conference on Computer Vision*, 2018.
[22] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[23] Y. H. Chen, W. Li, and L. V. Gool, "ROAD: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7892-7901.
[24] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472-7481.
[25] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. International Conference on Machine Learning*, 2018.
[26] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4500-4509.
[27] J. Hoffman, D. Q. Wang, F. Yu, and T. Darrell, "FCNs in the Wild: Pixel-level adversarial and constraint-based adaptation," arXiv preprint arXiv:1612.02649, 2016
[28] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 1992-2001.
[29] Y. H. Chen, W. Li, X. R. Chen, and L. V. Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1841-1850.
[30] I. J. Goodfellow, M. Mirza, B. Xu *et al.*, "Generative adversarial nets," in *Proc. Conference and Workshop on Neural Information Processing Systems*, 2014.
[31] A.Krizhevsky, L. Pearlstein *et al.*, "Investigating GAN and VAE to train DCNN," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, pp. 774-781, December 2019.
[32] Y. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Conference and Workshop on Neural Information Processing Systems*, 2012.
[33] K. Simonyan and A. Zisserman, "Very deep convolution networks for large-scale image recognition," in *Proc. International Consciousness Research Laboratories*, 2015.
[34] C. Szegedy, W. Liu, P. Sermanet *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[35] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.
[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[37] T.-Y. Lin, M. Maire, S. Belongie *et al.*, "Microsoft COCO: Common objects in context," in *Proc. European Conference on Computer Vision*, 2014, pp. 740-755.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," arXiv preprint arXiv:1505.04597, 2015.
[39] P. Isola, J.-Y. Zhu, T. H. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125-1134.
[40] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. International Conference on Machine Learning*, 2013.
[41] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2016.
[42] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?" in *Proc. The British Machine Vision Conference*, 2013.

Huachen Yu is a Ph.D. student at Meijo University in mechanical engineering. He received the master of engineering from the Department of Mechatronic Engineering, Nanjing Tech University in 2016. Now his research interests are focusing on computer vision, deep learning, and mobile robot system.

Jianming Yang is currently a professor of the Department of Mechanical Engineering, Meijo University, Japan. He received his doctor's degree in engineering from Nagoya University in 1995. His research interests include signal processing, robot vision, mobile robot system.