Self-Ensembling Attention Networks: Addressing Domain Shift for Semantic Segmentation

Yonghao Xu,[1, 2] Bo Du,[1*] Lefei Zhang,[1] Qian Zhang,[3] Guoli Wang,[3] Liangpei Zhang[2]

[1]School of Computer Science, Wuhan University, Wuhan 430072, P. R. China.
[2]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing,
Wuhan University, Wuhan 430079, P. R. China.
[3]Horizon Robotics, Inc., Beijing 100190, P. R. China.
yonghaoxu@ieee.org, remoteking@whu.edu.cn, zhanglefei@whu.edu.cn,
qian01.zhang@horizon.ai, guoli.wang@horizon.ai, zlp62@whu.edu.cn

Abstract

Recent years have witnessed the great success of deep learning models in semantic segmentation. Nevertheless, these models may not generalize well to unseen image domains due to the phenomenon of domain shift. Since pixel-level annotations are laborious to collect, developing algorithms which can adapt labeled data from source domain to target domain is of great significance. To this end, we propose self-ensembling attention networks to reduce the domain gap between different datasets. To the best of our knowledge, the proposed method is the first attempt to introduce self-ensembling model to domain adaptation for semantic segmentation, which provides a different view on how to learn domain-invariant features. Besides, since different regions in the image usually correspond to different levels of domain gap, we introduce the attention mechanism into the proposed framework to generate attention-aware features, which are further utilized to guide the calculation of consistency loss in the target domain. Experiments on two benchmark datasets demonstrate that the proposed framework can yield competitive performance compared with the state of the art methods.

Introduction

Semantic segmentation is a fundamental task in computer vision, which assigns the class label for each pixel in a given image (Tao et al. 2017). It has been widely utilized in many important applications nowadays such as autonomous driving (Xu et al. 2017).

Deep learning is a powerful tool for semantic segmentation task. Nevertheless, the training of deep learning models generally relies on a large amount of labeled data. A common solution in real world application is manual labeling. Obviously, this process is laborious and time-consuming. It is reported that high-quality semantic labeling requires about 90 minutes per image for the CITYSCAPES dataset (Cordts et al. 2016). An alternative way is to train deep neural networks with synthetic data. Recent advances on the graphics engine make it feasible to collect photo-realistic images from computer games with accurate pixel-level annotations automatically (Richter et al. 2016). Research has been made to apply models trained on synthetic data to real ones. However, the performance of these models on real images is usually not satisfactory due to the phenomenon of domain shift (Adel, Zhao, and Wong 2017), which is caused by factors like imaging devices (graphics engines vs. RGB sensors), illumination, distortion, and shadows in different domains, though it may be mild to a human observer. As a result, models which achieve high accuracies in synthetic datasets may fail to yield good performances for real images as shown in Figure 1 (a).

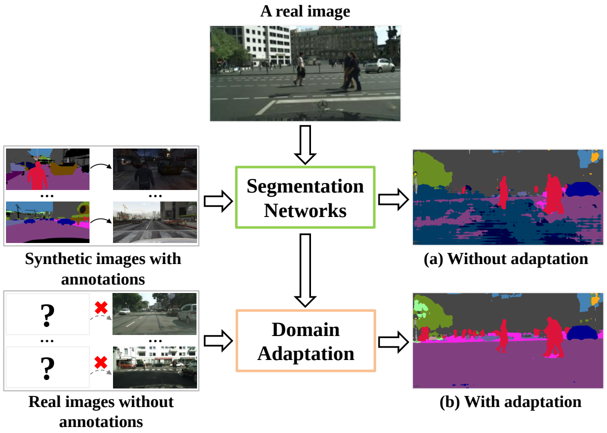

Figure 1: An illustration of domain shift in semantic segmentation task. (a) Segmentation map of a real image generated with models trained on synthetic data without domain adaptation technique. (b) Segmentation map of a real image with models trained on both synthetic data and real ones using domain adaptation technique.

*Corresponding author. This work is supported by the National Natural Science Foundation of China under Grant 61822113, Grant 41431175, Grant 61471274, and Grant U1536204.

An intuitive way to address this problem is to learn invariant features for images in both source domain and target domain and reduce the domain gap. This technique is also known as domain adaptation, which can be further divided into supervised one (Tzeng et al. 2014), unsupervised one (Adel, Zhao, and Wong 2017), and semi-supervised one (Ao, Li, and Ling 2017). Since the labels from the target domain are not available in most practical situations, this study focuses on unsupervised domain adaptation, where the source domain supplies both images and annotations and the target domain only supplies unlabeled images.

The initial inspiration of our work comes from an observation that a human can generally make a consistent interpretation without effort when an image changes slightly. We believe that a good model would possess this characteristic just like humans do. In the light of this statement, an ensemble of multiple networks is a better model than a single network, since the ensemble model generally yields better predictions and shows more robustness towards noise (Laine and Aila 2016; Singh, Hoiem, and Forsyth 2016). Given a series of networks trained on source-domain data, the ensemble prediction of these networks on target-domain images is also likely to be closer to the ground truth. Thus, a natural idea is to adopt the ensemble predictions as the pseudo labels to assist the training of the base networks in the ensemble (Bachman, Alsharif, and Precup 2014). As the iteration goes on, the base networks become more accurate, and the ensemble predictions also get closer to the correct labels in the target domain. In this way, the domain gap can be reduced correspondingly.

Obviously, the quality of the ensemble model determines the performance of the whole framework. Therefore, how to construct a good ensemble model is a critical issue. On the one hand, using more base networks in the ensemble model generally yields more robust results. On the other hand, it also brings about higher time costs in the training phase. To tackle this problem, we propose a novel self-ensembling attention network which not only inherits the advantage of ensemble model but also avoids high time costs for training. As shown in Figure 2, there are two main components: a student network which plays a role of base networks and a teacher network which plays a role of the ensemble networks. Both networks share the same architecture with an attention module inside. The student network is jointly optimized with supervised segmentation loss from the source domain and the unsupervised consistency loss from the target domain. The teacher network does not participate in the back-propagation, and is updated with an exponential moving average method using the parameters in the student network at different training time steps. In the test phase, the target-domain images are sent to the teacher network to obtain domain-invariant segmentation maps as shown in Figure 1 (b).

The main contributions of this study are summarized as follows:

(1) We propose a self-ensembling model to address domain shift in semantic segmentation task for the first time. Our solution provides a different view on how to learn domain-invariant features for semantic segmentation.

(2) Different regions in the image usually correspond to different levels of domain gap. Thus, those noteworthy regions deserve more attention. To this end, we introduce the attention mechanism into the proposed framework to generate attention-aware features. The learnt attention maps are further utilized to guide the calculation of consistency loss in the target domain which improves the performance of the model.

(3) Experiments on two challenging benchmark datasets demonstrate that our method significantly outperforms the state of the art domain adaptation methods.

Related Work

Semantic Segmentation

Different from traditional image classification task where each image is labeled with only one label, semantic segmentation requires pixel-level predictions, which is more challenging. Inspired by the work in (Long, Shelhamer, and Darrell 2015), numerous deep models have been proposed to tackle semantic segmentation task with fully convolutional networks (Noh, Hong, and Han 2015; Chen et al. 2018). To train these deep models, abundant pixel-level annotations are usually required, which are hard to be collected in real world applications (Tao et al. 2017).

An alternative approach is to train these deep models with synthetic data. Recent researches have made it feasible to generate dense pixel-accurate semantic label maps for photo-realistic images extracted from computer games automatically. It is reported that the labeling process for 25 thousand images obtained from the game Grand Theft Auto V costs only 49 hours, which dramatically reduces the amount of human effort required (Richter et al. 2016). However, due to the phenomenon of domain shift, models trained on these synthetic data can hardly yield satisfactory performance in real scenarios (Chang et al. 2017).

Domain Adaptation

The precondition of most conventional machine learning algorithms lies on the consistent underlying distribution of training and test sets (Liu, Wang, and Qiao 2017). Nevertheless, in the real-world applications, there usually exists discrepancy between training and test phases, resulting in poor performances (Killian et al. 2017). Domain adaptation is a class of techniques that aims to reduce the discrepancy between different domains (Tan et al. 2017).

According to whether labeled data in the target domain are available, domain adaptation can be divided into supervised one, unsupervised one, and semi-supervised one. Tzeng et al. propose a new CNN architecture to address the supervised domain adaptation which introduces an adaptation layer and an additional domain confusion loss to learn a representation that is both semantically meaningful and domain-invariant (Tzeng et al. 2014). When there is no labeled data in the target domain, the task is known as unsupervised domain adaptation. The framework proposed in

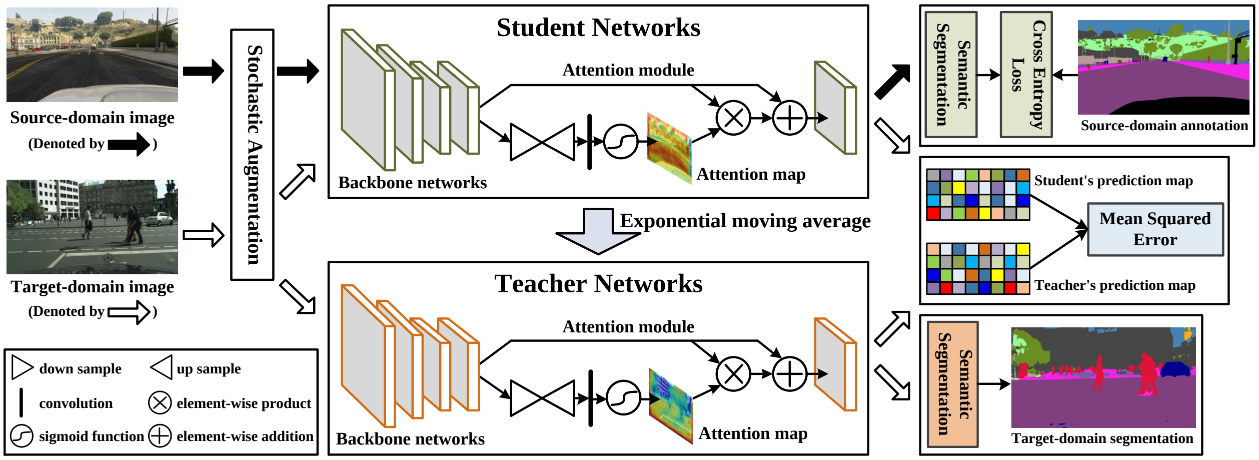

Figure 2: An illustration of the proposed self-ensembling attention networks. It consists of two main components: a student network and a teacher network. Both networks share the same architecture with an attention module inside. The source-domain images (denoted by solid arrows) are only input to the student network to calculate the segmentation loss. The target-domain images (denoted by hollow arrows) are input to both student and teacher networks to calculate the consistency loss. Then, the student network is jointly optimized with supervised segmentation loss from the source domain and the unsupervised consistency loss from the target domain. Notice that the teacher network does not participate in the back-propagation, and is updated with an exponential moving average method. In the test phase, the target-domain images are sent to the teacher network to accomplish the semantic segmentation.

(Tzeng et al. 2017) is a representative work where adversarial learning is utilized to improve the generalization ability of the model. By contrast, if both labeled and unlabeled data in the target domain are available, we refer it to semi-supervised domain adaptation. Long et al. propose a deep adaptation network to deal with this situation. The hidden representations of all task-specific layers are embedded in a reproducing kernel Hilbert space where the mean embedding of different domain distributions can be explicitly matched (Long et al. 2015).

Domain Adaptation for Semantic Segmentation

So far, most of the existing domain adaptation methods are designed for image classification task and seldom of them aim to address semantic segmentation task (Tzeng et al. 2014; 2017; Long et al. 2015). Hoffman et al. pioneer this research area with global and category specific adaptation techniques using pixel-level adversarial training (Hoffman et al. 2016). Another related work is curriculum-style learning where the curriculum domain adaptation solves easy tasks first to infer necessary properties about the target domain, such as label distributions over images and local distributions over landmark super-pixels. Then a segmentation network is trained with regularization that its predictions in the target domain follow those inferred properties (Zhang, David, and Gong 2017).

While the aforementioned domain adaptation methods mainly utilize adversarial training to reduce the domain gap, we propose a self-ensembling model to address this problem, which provides a different viewpoint on how to learn domain-invariant features for semantic segmentation. Self-ensembling model has achieved excellent results in semi-supervised learning (Laine and Aila 2016; Tarvainen and Valpola 2017). French et al. extend this model to deal with the unsupervised domain adaptation (French, Mackiewicz, and Fisher 2018), but this work does not propose any segmentation specific adaptation approach. To the best of the authors' knowledge, our method is the first attempt to introduce self-ensembling model to domain adaptation for semantic segmentation.

Self-Ensembling Attention Networks

In this section, we present the proposed self-ensembling attention networks for domain adaptation in detail.

Overview of the Proposed Model

As shown in Figure 2, there are two main components in the proposed model: a student network which plays a role of base networks and a teacher network which plays a role of the ensemble networks. Both networks share a consistent architecture with an attention module inside. In our implementations, we employ the DeepLab-v2 (Chen et al. 2018) with VGG-16 (Simonyan and Zisserman 2014) model pre-trained on ImageNet (Deng et al. 2009) as the backbone networks. Stochastic augmentation is implemented for images in both source and target domains to increase the generalization ability of the model (French, Mackiewicz, and Fisher 2018). More specifically, we add Gaussian noise with a mean of zero and a standard deviation of 0.1 to each pixel in the image. The source-domain images are only fed into the student network to calculate the segmentation loss. The target-domain images are input to both student and teacher networks to calculate the consistency loss. The student network is then optimized with the combination of the segmentation loss and the consistency loss. By contrast, the teacher network does not participate in the back-propagation, and is updated with an exponential moving average method using the parameters in the student network at different time steps.

Owing to the regularization of the consistency loss, the student network can thereby learn from the output of the teacher network, which is likely to be closer to the ground truth in the target domain. As the iteration goes on, the student network becomes more accurate, and the ensemble predictions in the teacher network also get closer to the correct labels in the target domain. In this way, domain-invariant features can be learnt correspondingly. In the test phase, the target-domain images are sent to the teacher network to accomplish the semantic segmentation.

Attention Module

Different regions in the image usually correspond to different levels of domain gap. Thus, those noteworthy regions deserve more attention. Researches on human perception process demonstrate the significance of attention mechanism, which uses the high-level information to guide bottom-up feedforward process (Mnih et al. 2014). Inspired by the work in (Chen et al. 2016; Wang et al. 2017; Zhang et al. 2018), we introduce the attention module into the domain adaptation framework to learn attention-aware features.

The architecture of the proposed attention module is simple and straightforward. Let $F(x)$ be the output of the backbone networks with an input data x. The attention map $A(x)$ is formulated as:

$$A(x) = T(U(D(F(x)))) \tag{1}$$

where $D(\cdot)$ represents the down-sample operation with 2×2 average pooling, $U(\cdot)$ represents the up-sample operation with bilinear interpolation, and $T(\cdot)$ represents the nonlinear transformation with an 1×1 convolutional layer and sigmoid activation. Then, the output of the attention module $H(x)$ can be formulated as:

$$H(x) = (1 + A(x)) * F(x) \tag{2}$$

where $*$ denotes the element-wise product. Since the attention map $A(x)$ ranges from $[0, 1]$, it acts as control gates for features in $F(x)$. If $A(x) = 0$ for all positions in the feature map, then, the attention feature $H(x)$ degenerates into the original $F(x)$. Otherwise, features in $F(x)$ can get heightened in some positions and restrained in other positions. In the next subsection, we will take advantage of this property to assist the computation of the consistency loss.

Optimization

Given an image X_s and a corresponding label map Y_s in the source domain, we first define the segmentation loss with cross-entropy as:

$$\mathcal{L}_{seg}(X_s) = -\frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{W}\sum_{c=1}^{C}\Big[Y_s^{(u,v,c)} \cdot \log\Big(P_S(g(X_s))^{(u,v,c)}\Big)\Big] \tag{3}$$

where P_S denotes the probability map generated by the student network and $g(\cdot)$ denotes the stochastic augmentation. H , W , and C represent the height, width, and number of categories, respectively.

In order to learn domain-invariant features, we then define the consistency loss with mean squared error as:

$$\mathcal{L}_{con}(X_t) = -\frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{W}\sum_{c=1}^{C}\Big[M^{(u,v)} \cdot \Big\|P_S(g(X_t))^{(u,v,c)} - P_T(g(X_t))^{(u,v,c)}\Big\|^2\Big] \tag{4}$$

where P_T denotes the probability map generated by the teacher network and X_t denotes the target-domain image. $M \in \mathbb{Z}^{H \times W}$ is the attention mask matrix which can be defined as:

$$M^{(u,v)} = \begin{cases} 1, \text{if } A_T^{(u,v)} > \tau_{att} \\ 0, \text{otherwise} \end{cases} \tag{5}$$

where τ_{att} denotes the attention threshold and A_T is the attention map generated by the teacher network. In this way, only those regions that have higher attention activation than τ_{att} can participate in the calculation of consistency loss. Then, the overall loss function of the student network can be formulated as:

$$\mathcal{L}_S = \mathcal{L}_{seg}(X_s) + \lambda_{con}\mathcal{L}_{con}(X_t) \tag{6}$$

where λ_{con} is a weighting factor for the unsupervised consistency loss. Notice that the teacher network does not participate in the back-propagation since it acts as an ensemble model. Instead, we utilize the exponential moving average method to update the parameters in the teacher network. Let θ_T^{t-1} be the parameters in the teacher network at $(t-1)$th iteration. Then, at the tth iteration, θ_T^t can be calculated as:

$$\theta_T^t = \alpha\theta_T^{t-1} + (1-\alpha)\theta_S^t \tag{7}$$

where θ_S^t denotes the parameters in the student network at tth iteration and α is a smoothing coefficient hyper-parameter.

Experiments

In this section, we first introduce the datasets utilized in this study. Then, the experimental results are presented and analyzed in detail.

Datasets

We use the CITYSCAPES (Cordts et al. 2016) as our target-domain data in the experiments. For the source domain, two challenging synthetic datasets including SYNTHIA (Ros et al. 2016) and GTA-5 (Richter et al. 2016) are utilized.

CITYSCAPES is a real-world vehicle-egocentric image dataset collected from 50 cities in Germany and the countries around. It provides three disjoint subsets: 2975 training images, 500 validation images, and 1525 test images. It also provides accurate pixel-level annotations for all images with 19 different categories. In order to ensure the fairness of experimental results, we follow the same evaluation protocol

Table 1: Results of semantic segmentation by adapting from SYTNHIA to CITYSCAPES. MCD (Saito et al. 2018) and CyCADA (Hoffman et al. 2018) do not report the experimental results on GTA-5 dataset with VGG-16 backbone networks. Thus we omit them in this table. The IoUs of wall, fence, and pole in CCA are not reported (Chen et al. 2017). For the remaining 13 classes, the mean IoU of CCA is 35.7%, while our method achieves 43.6% in this case.

SYNTHIA→CITYSCAPES																	
Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	sky	person	rider	car	bus	motocycle	bicycle	mIoU (%)
NoAdapt	6.4	17.7	29.7	1.2	0.0	15.1	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	17.4
FCN Wld	11.5	19.6	30.8	**4.4**	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.2
CDA	65.2	26.1	74.9	0.1	**0.5**	10.7	3.5	3.0	76.1	70.6	47.1	8.2	43.2	**20.7**	0.7	13.1	29.0
CCA	62.7	25.6	78.3	-	-	-	1.2	5.4	81.3	**81.0**	37.4	6.4	63.5	16.1	1.2	4.6	35.7
Ours	**86.8**	**39.2**	**79.2**	2.7	0.3	**29.3**	**3.6**	**14.7**	**81.5**	78.7	**52.9**	**11.4**	**79.7**	18.5	**5.9**	**15.2**	**37.5**

as specified by the previous works (Hoffman et al. 2016; Zhang, David, and Gong 2017). In the test phase, we evaluate on the CITYSCAPES validation set with 500 images.

SYNTHIA is a large dataset of photo-realistic frames rendered from a virtual city with precise pixel-level annotations. Following previous works (Hoffman et al. 2016; Zhang, David, and Gong 2017), we use the SYNTHIA-RAND-CITYSCAPES subset that contains 9400 images with annotations which are compatible with CITYSCAPES dataset. The 16 common categories between SYNTHIA and CITYSCAPES are selected to make quantitative assessment. These classes are: road, sidewalk, building, wall, fence, pole, light, sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle.

GTA-5 contains 24966 high quality labeled frames from realistic open-world computer games, Grand Theft Auto V (GTA-5). Each frame is generated from fictional city of Los Santos, based on Los Angeles in Southern California with annotations that are compatible with CITYSCAPES dataset. We use all the 19 official training classes in our experiment including road, sidewalk, building, wall, fence, pole, light, sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle.

Implementation Details

In our implementations, we employ the DeepLab-v2 (Chen et al. 2018) with VGG-16 (Simonyan and Zisserman 2014) model pre-trained on ImageNet (Deng et al. 2009) and Pascal VOC datasets (Everingham et al. 2010) as the backbone networks. The Adam optimizer (Kinga and Adam 2015) with a learning rate of $1e-5$ and weight decay of $5e-5$ is utilized to train the proposed networks. Each mini-batch consists of 1 source-domain image and 1 target-domain image. We resize all the images to the size of 1024×512. While evaluating on CITYSCAPES dataset whose images and ground truth annotations have a size of 2048×1024, we first produce our predictions on the 1024×512 sized image and then up-sample our predictions by a factor of 2 to get the final label map, which is used for evaluation. The experiments in this paper are implemented in PyTorch with a single NVIDIA GTX TITAN X GPU.

Performance Evaluation

In this subsection, we report the semantic segmentation results of the proposed method along with the state of the art methods. To ensure the fairness of the comparison, all methods reported here utilize the VGG-16 as the backbone networks. A brief introduction about the comparing methods are given as below.

No adaptation (NoAdapt)(Hoffman et al. 2016) directly trains the segmentation networks on SYNTHIA and GTA-5 without any domain adaptation, which is the baseline for the experiments.

FCNs in the wild (FCN Wld) (Hoffman et al. 2016) introduces a pixel-level adversarial loss to the intermediate layers of the network and impose constraints on label statistics to the network output.

Curriculum domain adaptation (CDA, in ICCV 2017) (Zhang, David, and Gong 2017) proposes a curriculum-style learning approach to minimize the domain gap in semantic segmentation. The curriculum domain adaptation first solves easy tasks such as estimating label distributions, then infers the necessary properties about the target domain.

Cross city adaptation (CCA, in ICCV 2017) (Chen et al. 2017) advances a joint global and class-specific domain adversarial learning framework. Adaptation of pre-trained segmentation networks to other domains can be achieved via this framework without the need of any user annotation or interaction.

Maximum classifier discrepancy (MCD, in CVPR 2018) (Saito et al. 2018) learns two classifiers from the source domain and maximizes their disagreement on the target images in order to detect target examples that fall out of the support of the source domain. After that, it updates the generator to minimize the two classifiers' disagreement on the target domain.

Cycle-consistent adversarial domain adaptation (CyCADA, in ICML 2018) (Hoffman et al. 2018) transforms the synthetic images of the source domain to the style of the target domain (real images) using CycleGAN (Zhu et al. 2017).

As shown in Table 1 and Table 2, all domain adaptation results are significantly better than those without adaptation (NoAdapt), which demonstrates the large domain gap between synthetic images and real ones. In both datasets,

Table 2: Results of semantic segmentation by adapting from GTA-5 to CITYSCAPES. CCA (Chen et al. 2017) does not report the experimental results on GTA-5 dataset with VGG-16 backbone networks. Thus we omit it in this table.

GTA-5→CITYSCAPES																				
Methods	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bicycle	mIoU (%)
NoAdapt	31.9	18.9	47.7	7.4	3.1	16.0	10.4	1.0	76.5	13.0	58.9	36.0	1.0	67.1	9.5	3.7	0.0	0.0	0.0	21.2
FCN Wld	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
CDA	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	**16.6**	28.9
MCD	**86.4**	8.5	76.1	18.6	9.7	14.9	7.8	0.6	**82.8**	**32.7**	71.4	25.2	1.1	76.3	16.1	17.1	1.4	0.2	0.0	28.8
CyCADA	83.5	**38.3**	76.4	20.6	**16.5**	22.2	26.2	**21.9**	80.4	28.7	65.7	49.4	4.2	74.6	16.0	**26.6**	**2.0**	8.0	0.0	34.8
Ours	85.2	19.8	**77.8**	**25.0**	12.2	**28.7**	**29.8**	13.6	79.2	25.9	**74.9**	**54.1**	**12.1**	**82.7**	**21.6**	10.4	0.0	**19.9**	5.1	**35.7**

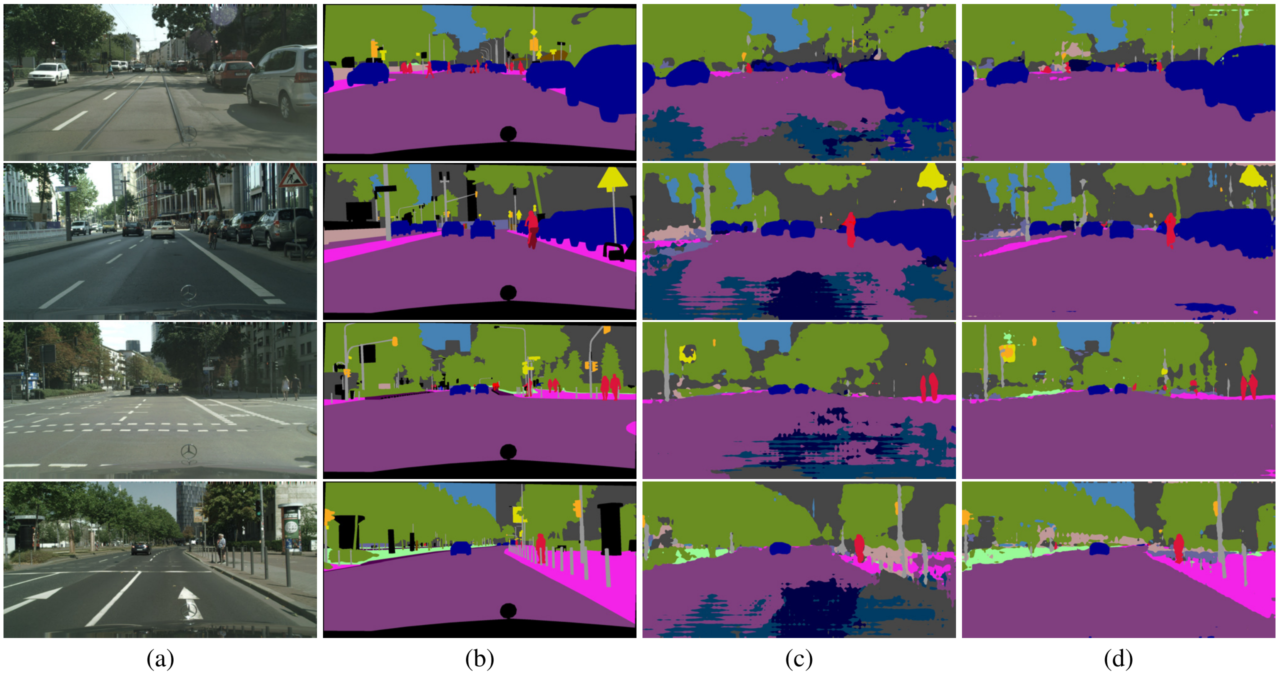

Figure 3: Qualitative Results. (a) Input images from the CITYSCAPES dataset. (b) Ground-truth annotations. (c) Segmentation maps with models trained on the GTA-5 dataset without domain adaptation technique. (d) Testing results of the model adapted from GTA-5 dataset.

our proposed method achieves higher class-wise IoU than the NoAdapt baseline for any class. The mean IoUs of our method get about 20% and 14% higher than those of the NoAdapt baseline in SYTNHIA and GTA-5 datasets, respectively. These results verify the effectiveness of the proposed method for learning domain-invariant features. Compared with the state of the art approaches (Hoffman et al. 2016; Zhang, David, and Gong 2017; Chen et al. 2017; Saito et al. 2018; Hoffman et al. 2018), our method also outperforms them by a large margin. Note that the IoUs of wall, fence, and pole in CCA (Chen et al. 2017) are not reported. For the remaining 13 classes, the mIoU of CCA is 35.7%, while our method achieves 47.7% in this case. The qualitative results are shown in Figure 3.

Parameter Analysis

The attention threshold τ_{att} and the weighting factor λ_{con} for the unsupervised consistency loss are two important parameters. In this subsection, we evaluate the performance of the proposed framework with different τ_{att} and λ_{con} on SYTNHIA dataset. The smoothing coefficient α in the exponential moving average is empirically set as 0.99 in our experiments.

The weighting factor λ_{con} controls the balance between the unsupervised consistency loss and the supervised segmentation loss. On the one hand, setting a too small λ_{con} would make the framework ignore the adaptation part in the target-domain data during the training. On the other hand, a large λ_{con} would prevent the framework from learning good representations on the source-domain data, making the student network very poor. The experimental results in Table 3 demonstrate that a larger λ_{con} is more detrimental than a smaller one. In the case of $\lambda_{con} = 10$, the mIoU is only 21.4, which is much worse than the case of small λ_{con}.

As for the attention threshold τ_{att}, generally speaking, adaptation with high attention threshold fails to yield good mIoUs since the attention mask filters most of the regions in the image. Thus, very few pixels can participate in the calculation of consistency loss, resulting in poor performance. As shown in Table 4, in the case of $\tau_{att} = 0.5$, the mIoU is only 31.2, which is even worse than $\tau_{att} = 0$ case. By contrast, choosing a relatively small threshold value like 0.3 can enable the framework to concentrate on the adaptation of those noteworthy regions.

Table 3: Parameter analysis of the weighting factor λ_{con} for the unsupervised consistency loss.

SYTNHIA→CITYSCAPES					
λ_{con}	0.1	0.3	1	3	10
mIoU (%)	36.3	37.5	36.1	24.2	21.4

Table 4: Parameter analysis of the attention threshold τ_{att}.

SYTNHIA→CITYSCAPES					
τ_{att}	0.0	0.1	0.3	0.5	0.7
mIoU (%)	34.6	36.2	37.5	31.2	31.9

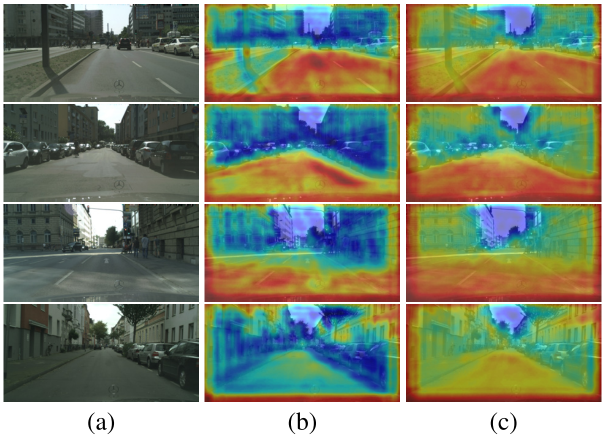

Figure 4: Illustrations of the learnt attention maps. (a) Input images from the CITYSCAPES dataset. (b) Attention maps obtained at the 500th iteration. (c) Attention maps obtained at the 1500th iteration. Red regions in the map correspond to high attention while blue ones correspond to low attention.

In order to further investigate where the attention network puts more attention in the domain adaptation, we further visualize the learnt attention maps. As shown in Figure 4 (b), during the early period of the training, the network mainly concentrates on the road regions in the image. As the iteration goes on, the attention expands to wider regions, but the road category is still the most attractive one, as shown in Figure 4 (c). It can also be observed that the sky regions attract the least attention from the network. One reason for this phenomenon may lie on the fact that the sky objects share a relatively similar appearance in both source and target domains. According to the results in Table 1, even without domain adaptation, the segmentation network directly trained on source domain can get an IoU of 66.8% for the sky class in target domain, which is a very high score. By contrast, the NoAdapt method can only get an IoU of 6.4% for the road class in target domain. Therefore, the proposed attention mechanism does help the framework to focus more on those noteworthy regions.

Conclusion

In this paper, we propose self-ensembling attention networks to address domain shift for semantic segmentation. Considering that different regions in the image usually correspond to different levels of domain gap, we introduce the attention mechanism into the proposed framework to generate attention-aware features, which are further utilized to guide the calculation of consistency loss in the target domain. There are two main components in the proposed framework: a student network which plays a role of base networks and a teacher network which plays a role of the ensemble networks. With the help of the consistency loss, the student network can thereby learn from the output of the teacher network. As the iteration goes on, the student network becomes more accurate, and the ensemble predictions in the teacher network also get closer to the correct labels in the target domain. In this way, domain-invariant features can be learnt correspondingly. Experiments on two benchmark datasets demonstrate that the proposed framework can yield competitive performance compared with the state of the art methods.

Since the performance of the framework depends largely on the quality of teacher-generated predictions, our future work will try to further improve the robustness of the teacher network.

References

Adel, T.; Zhao, H.; and Wong, A. 2017. Unsupervised domain adaptation with a relaxed covariate shift assumption. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Ao, S.; Li, X.; and Ling, C. X. 2017. Fast generalized distillation for semi-supervised domain adaptation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. In *International Conference on Neural Information Processing Systems (NIPS)*, 3365–3373.

Chang, W. C.; Wu, Y.; Liu, H.; and Yang, Y. 2017. Cross-domain kernel induction for transfer learning. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016. Attention to scale: Scale-aware semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3640–3649.

Chen, Y.-H.; Chen, W.-Y.; Chen, Y.-T.; Tsai, B.-C.; Wang, Y.-C. F.; and Sun, M. 2017. No more discrimination: Cross city adaptation of road scene segmenters. In *IEEE International Conference on Computer Vision (ICCV)*, 2011–2020.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4):834–848.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–338.

French, G.; Mackiewicz, M.; and Fisher, M. 2018. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations (ICLR)*.

Hoffman, J.; Wang, D.; Yu, F.; and Darrell, T. 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*.

Killian, T.; Daulton, S.; Konidaris, G.; and Doshivelez, F. 2017. Robust and efficient transfer learning with hidden-parameter markov decision processes. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Kinga, D., and Adam, J. B. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5.

Laine, S., and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint*.

Liu, J.; Wang, Y.; and Qiao, Y. 2017. Sparse deep transfer learning for convolutional neural network. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *International Conference on Neural Information Processing Systems (NIPS)*, 2204–2212.

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 1520–1528.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, 102–118. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3234–3243.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, S.; Hoiem, D.; and Forsyth, D. 2016. Swapout: Learning an ensemble of deep architectures. In *International Conference on Neural Information Processing Systems (NIPS)*.

Tan, B.; Zhang, Y.; Pan, S. J.; and Yang, Q. 2017. Distant domain transfer learning. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Tao, Z.; Liu, H.; Fu, H.; and Fu, Y. 2017. Image cosegmentation via saliency-guided constrained clustering with cosine similarity. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Tarvainen, A., and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *International Conference on Neural Information Processing Systems (NIPS)*, 1195–1204.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*.

Xu, H.; Gao, Y.; Yu, F.; and Darrell, T. 2017. End-to-end learning of driving models from large-scale video datasets. *arXiv preprint*.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Y.; David, P.; and Gong, B. 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In *IEEE International Conference on Computer Vision (ICCV)*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*.