Cross-Domain Semantic Segmentation via Domain-Invariant Interactive Relation Transfer

Fengmao Lv[1*†] Tao Liang[1,2†] Xiang Chen[2] Guosheng Lin[3]

[1]Center of Statistical Research & School of Statistics, Southwestern University of Finance and Economics, China
[2]Content Platform of Center & Content Development Platform, Tencent, China
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
{fengmaolv,taoliangdpg}@126.com chx245296678@gmail.com gslin@ntu.edu.sg

Abstract

Exploiting photo-realistic synthetic data to train semantic segmentation models has received increasing attention over the past years. However, the domain mismatch between synthetic and real images will cause a significant performance drop when the model trained with synthetic images is directly applied to real-world scenarios. In this paper, we propose a new domain adaptation approach, called Pivot Interaction Transfer (PIT). Our method mainly focuses on constructing pivot information that is common knowledge shared across domains as a bridge to promote the adaptation of semantic segmentation model from synthetic domains to real-world domains. Specifically, we first infer the image-level category information about the target images, which is then utilized to facilitate pixel-level transfer for semantic segmentation, with the assumption that the interactive relation between the image-level category information and the pixel-level semantic information is invariant across domains. To this end, we propose a novel multi-level region expansion mechanism that aligns both the image-level and pixel-level information. Comprehensive experiments on the adaptation from both GTAV and SYNTHIA to Cityscapes clearly demonstrate the superiority of our method.

1. Introduction

Towards thorough autonomous-driving, road scene segmentation, which aims to acquire detailed understanding about road environments, is a key functionality. Over the past few years, deep convolutional neural networks have gained great advances in semantic segmentation [21, 2, 19]. However, to train a good segmentation model, it usually requires a huge amount of time and labor effort to obtain sufficient pixel-level annotations of real-world images beforehand. To reduce the labeling consumption, one can turn to collect photo-realistic synthetic data from game simulators (e.g., Grand Theft Auto V), where the full pixel-level annotations can be generated automatically [27, 28]. But due to the clear domain difference between the synthetic (source) images and real-world (target) images, it will cause a dramatic performance drop if the model trained with the synthetic images is directly applied to real-world scenarios.

In order to promote the adaptation of semantic segmentation model from synthetic domains to real-world domains, we propose a novel domain adaptation approach dubbed **P**ivot **I**nteraction **T**ransfer (PIT) by constructing pivot information that is common knowledge shared across domains to bridge the domain gap. To this end, we first infer the image-level category information about the target images, which can be less challenging to estimate due to the strong idiosyncrasies of urban traffic scenes (e.g., the region size and spatial layout of buildings, lights, persons, etc.) [41]. There exists a strong interaction between the image-level categories and the pixel-level semantic information: the image-level category information provides global estimation of the categories appeared in the whole image, and in the meanwhile, suppresses implausible pixel-level label prediction for segmentation; on the other hand, the specific pixel-level label information in each image will immediately reveal the image-level categories. Since the images of urban traffic scenes have very strong idiosyncrasies [41], the interactive relation between the image-level categories and the pixel-level semantic information can be invariant

* Corresponding author: F. Lv (email: fengmaolv@126.com).
† indicates equal contribution.

regardless of different domains. Hence, we can treat this domain-invariant information as the transferrable factor that could be utilized to facilitate pixel-level transfer for semantic segmentation. Unlike the prior works that focus on learning domain-invariant representations of instances by using domain adversarial training [15, 32, 42, 8, 30], our method constructs the transferrable factors across different domains explicitly.

Our model includes two interactive components: 1) the **fine-grained** component produces the segmentation masks for semantic segmentation; 2) the **coarse-grained** component generates class activation maps through learning from the image-level category information. As displayed in Fig. 1, the coarse-grained component contains multiple region expansion units, each of which can produce a different class activation map by using a specific aggregation operation. To model the interactive relation between the two components, we adopt a reconstruction mechanism to reconstruct the class activation map of the fine-grained component by the class activation maps of the coarse-grained component. The knowledge transfer insight lies in sharing the reconstruction module between the source and target domains, with the intuition that the interactive relations between the category information of the two different granularity levels are domain-invariant. To this end, the reconstruction loss in the target domain promotes effective information flow from the inferred image-level category information to the pixel-level prediction. Unlike the prior works [42, 8, 34], our method does not involve the minimax games in adversarial training which are usually hard to solve.

The main contributions of this work are as follows:

- We propose a new domain adaptation approach by constructing pivot information that is common knowledge shared across domains to promote the adaptation of semantic segmentation models.
- We propose a multi-level region expansion mechanism to model the domain-invariant interaction between the image-level categories and the pixel-level semantic information.
- Comprehensive experiments on the adaptation from both GTAV and SYNTHIA to Cityscapes demonstrate that our method can obtain better results than the existing state-of-the-art works.

2. Related Works

2.1. Semantic segmentation

Semantic segmentation is a very important visual task. Based on Fully Convolutional Networks (FCN) [21], diverse advanced designs, e.g. multi-scale aggregation [2, 19, 40], context relation [44, 20], etc, are proposed to improve the performance of semantic segmentation neural networks. In addition, post-processing techniques such as conditional random fields [45] can also be utilized as an effective tool to further improve the segmentation performance. However, the training of segmentation networks requires a large amount of real-world images with pixel-level annotations, which are usually difficult to collect.

2.2. Domain adaptation

Machine learning works based on the assumption that the training and test data are drawn from the identical distribution. It will cause a dramatic performance drop if the training and test datasets have clear domain discrepancy. Domain adaptation focuses on how to bridge the domain gap and obtain good generalization performance on the target data. Over the past years, domain adaptation is mainly studied in image recognition [12, 43, 4, 17]. In general, the main idea of the existing domain adaptation methods is to learn domain-invariant representations in deep neural networks [22, 12, 43] or generate pseudo labels of target images to adapt the prediction models [29, 11].

2.3. Domain adaptation for semantic segmentation

Exploiting photo-realistic synthetic data to train semantic segmentation models has received increasing attention over the recent years. Similar to classification, one of the primary approaches towards domain adaptation for semantic segmentation is to learn domain-invariant representation of instances by domain adversarial training [8, 6, 30, 42, 10, 35, 23]. Alternative approaches are also explored. To be specific, the domain gap can also be reduced through directly translating the source images into target-style images with Generative Adversarial Networks (GANs) [25, 14, 7, 18]. Recently, the structural consistency between domains is exploited to promote the adaptation of semantic segmentation neural networks [34, 5, 1, 24, 37]. Additionally, domain adaptation can be implemented by learning from a pre-defined curriculum as well [31, 41]. Some other works are inspired by the methods in semi-supervised learning, such as self-training [47] or entropy minimization [3, 36].

3. Motivation

Pivot knowledge. With only sufficient pixel-level labels in the source domain, the distribution mismatch between domains makes it infeasible to directly apply a model trained with synthetic images to real-world images. Following the domain adaptation methods in image classification, the prior works propose to learn domain-invariant feature space of instances by using domain adversarial training [15, 32, 42, 8, 30]. However, this is not suitable for semantic segmentation since it remains unclear what comprises the data instances, i.e. the proxy to evaluate domain discrepancy, in semantic segmentation [41].

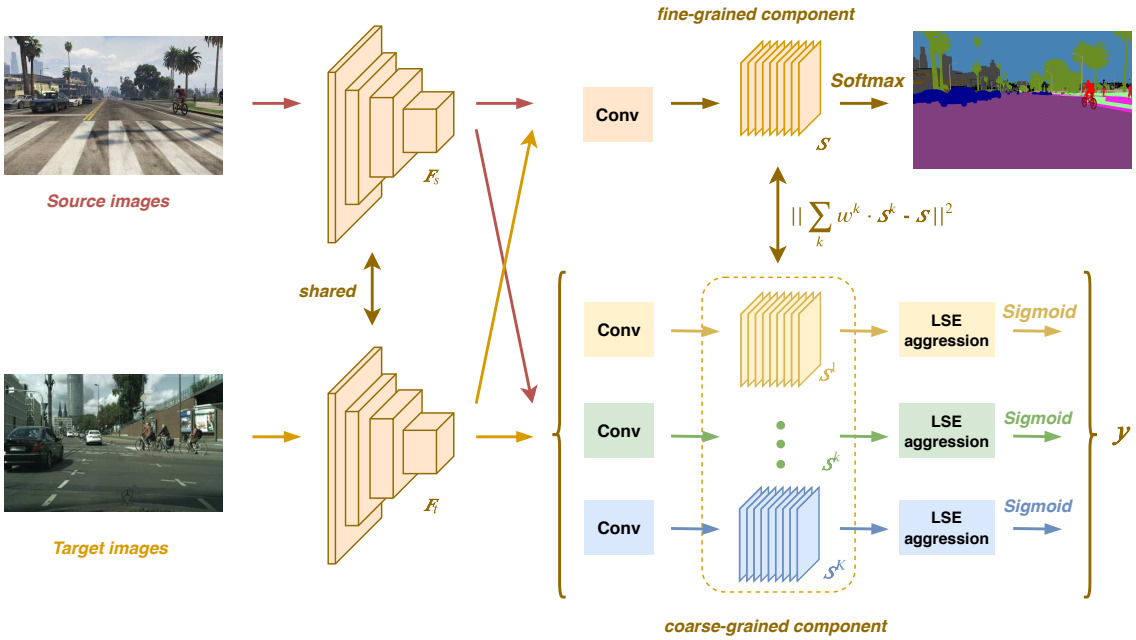


Figure 1: The overall architectural of our proposed model. The images from source and target domains will be input to both the fine-grained and coarse-grained components. Our method addresses the adaptation of semantic segmentation neural networks through modeling domain-invariant interaction between the fined-grained component using pixel-level category information and the coarse-grained component using image-level category information. In the corse-grained component, the region expansion units are specified by different colors. The notation "Conv" denotes a convolutional layer.

Unlike the prior methods, this work aims to address the adaptation of semantic segmentation neural networks by explicitly constructing pivot information that is invariant across different domains. Thanks to the strong idiosyncrasies of urban traffic scenes, it is less challenging to infer the image-level category information of target images than producing the pixel-level predictions [41]. Immediately, it is clear that there exists a strong association between image-level categories and pixel-level semantic information [46, 16]. We argue that the interactive relation between the category information of the two different granularity levels is invariant across domains. This assumption can be reasonable since the images of urban traffic scenes share many similarities regardless of different domains. For example, as shown in Fig. 1, the synthetic and real-world images have very similar appearance in both spatial layout and local context for one specific category, e.g. road, car, building, etc. Motivated by this, we consider the interaction between the image-level categories and the pixel-level semantic information as the pivot information for domain adaptation.

Multi-level region expansion transfer. A crucial point lies in how to model this interaction. To this end, our model includes two components to learn from the category information of different granularity levels, as shown in Fig. 1. The fine-grained component produces the segmentation masks for semantic segmentation, while the coarse-grained component generates class activation maps through learning from the image-level category information. The key insignt in learning from image-level information is to localize a specific category and expand its region to coincide with the boundary. However, for urban traffic scenes, it is usually difficult to determine the extent of region expansion since the region size varies a lot across different categories. Motivated by this, we propose to design multiple region expansion units in the coarse-grained component, each of which can produce a different class activation map by using a specific aggregation operation. The class activation map of each region expansion unit can reveal some aspect of the pixel-level ground truth, but they are generally inaccurate.

To associate the fine-grained and coarse-grained components, we integrate the class activation maps of the coarse-grained component with a weighting scheme and enforce them to reconstruct the class activation map of the fine-grained component. The transfer insight is reflected by sharing this reconstruction module between the source and target domains. To this end, the reconstruction loss in the target domain promotes effective information flow from the inferred image-level category information to the pixel-level prediction. Hence, the interactive relation captured from the source images can be transferred to the target domain,

which helps to infer the pixel-level labels given the inferred image-level category information.

4. Pivot interaction transfer

4.1. Problem statement

Denote by $\boldsymbol{I}_s \in \mathbb{R}^{H\times W\times 3}$ a source image and $\boldsymbol{Y}_s \in \{0,1\}^{H\times W\times C}$ the corresponding pixel-level labels, where H, W and C are the height, width and category number, respectively. In the problem of domain adaptation for semantic segmentation, we are given a labeled source dataset $\mathcal{D}_{\mathcal{S}} = \{(\boldsymbol{I}_s, \boldsymbol{Y}_s)\}_{s\in\mathcal{S}}$ and an unlabeled target dataset $\mathcal{D}_{\mathcal{T}} = \{\boldsymbol{I}_t\}_{t\in\mathcal{T}}$ for training. Our goal is to learn a good segmentation network that can achieve desirable pixel-level prediction performance over the target domain.

4.2. Approach

Image-level category information. Following [41], we extract image features from Inception-ResNet-v2 [33] and use multinomial logistic regression to infer the image-level category information of target images. We train the logistic regression model using the soft label $\boldsymbol{y}_s \in [0,1]^C$, which is calculated as follows:

$$\boldsymbol{y}_s(c) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{Y}_s(i,j,c).$$

It is clear that $\boldsymbol{y}_s$ reflects the occupancy proportion of each category over an image. For a target image, we directly take the output of the multinomial logistic regression model as its image-level category information, which is denoted by $\boldsymbol{y}_t$.

Fine-grained component. As displayed in Fig. 1, our model includes two interactive components to learn from different granularity levels. Both the fine-grained and coarse-grained components share a common feature extractor, i.e. the fully convolutional backbone, which transforms the input images $\boldsymbol{X}_m$ to high-level semantic features $\boldsymbol{F}_m$, where $m \in \mathcal{S}\cup\mathcal{T}$. Then, the fine-grained component further transforms $\boldsymbol{F}_m$ via a convolution layer and a bi-linear interpolation layer to produce the class activation map $\boldsymbol{S}_m \in \mathbb{R}^{H\times W\times C}$. The final pixel-level labels are produced by feeding $\boldsymbol{S}_m$ to the softmax layer: $\hat{\boldsymbol{Y}}_m = \mathrm{softmax}(\boldsymbol{S}_m)$.

Coarse-grained component. The coarse-grained component includes K region expansion units, each of which transforms $\boldsymbol{F}_m$ via a convolution layer and a bi-linear interpolation layer to produce a class activation map $\boldsymbol{S}_m^k \in \mathbb{R}^{H\times W\times C}$. Then, a Log-Sum-Exp (LSE) aggregation layer [26] will process $\boldsymbol{S}_m^k$ to generate an image-level vector $\boldsymbol{s}^k \in \mathbb{R}^{C\times 1}$:

$$\boldsymbol{s}^k(c) = \log\left[\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\exp(r^k\cdot\boldsymbol{S}_m^k(i,j,c))\right]. \quad (1)$$

The final image-level labels are produced by feeding $\boldsymbol{s}_m^k$ to the sigmoid layer: $\hat{\boldsymbol{y}}_m^k = \mathrm{sigmoid}(\boldsymbol{s}_m^k)$.

Each region expansion unit is trained with the image-level category information independently. The aggregation layer can drive the network towards good pixel-level assignments over the class activation map $\boldsymbol{S}_m^k$. In order to produce class activation maps that can reveal different aspects of the pixel-level ground truth, each unit uses a different aggregation operation, which is reflected by the value of r^k in Eq. 1. In principle, the LSE aggregation operation acts as a smooth version of the max pooling aggregation. The hyper-parameter r^k is the smooth parameter that controls how smoothly the activation value of each pixel contributes to the aggregation output value $\boldsymbol{s}^k(c)$. Specifically, the aggregation operation with a large value of r^k tends to select the regions with large activation values to generate the output. Hence, the region expansion unit can only expand a small region over $\boldsymbol{S}_m^k$ for each image-level category. On the other hand, the one with a small value of r^k tends to equally consider every spatial position in the activation map to generate the output and hence makes the region expansion unit expand a large region over $\boldsymbol{S}_m^k$. The class activation map of each region expansion unit can reveal some aspect of the pixel-level ground truth, but none of them is absolutely accurate, since the region size varies a lot across different categories in the images of urban traffic scenes. With multiple LSE expansion units, network is able to learn to perform different level of expansion and adapt to the object regions.

Interaction transfer. The individual components for image-level and pixel-level prediction cannot exploit the interactive relations between the category information of different granularity levels. Hence, we propose a reconstruction mechanism to model the interactions between these two components. To this end, we integrate the class activation maps of the coarse-grained component with a weighting scheme and enforce them to reconstruct the class activation map of the fine-grained component, which is formulated as follows:

$$d_m = \left\|\sum_{k=1}^{K} w^k\cdot\boldsymbol{S}_m^k - \boldsymbol{S}_m\right\|^2,$$

where w^k are learnable parameters. The reconstruction mechanism can promote to exchange knowledge across the two components and capture the domain-invariant interaction between the image-level categories and the pixel-level semantic information.

Training. In unsupervised domain adaptation, only the source images have pixel-level labels. Hence, we train the fine-grained component with source images:

$$\mathcal{L}_f = \sum_{m\in\mathcal{S}}\mathcal{L}_{seg}(\boldsymbol{Y}_m, \hat{\boldsymbol{Y}}_m),$$

where $\mathcal{L}_{seg}$ is the pixel-wise cross-entropy loss. The coarse-grained component is trained with both source and target images:

$$\mathcal{L}_c = \sum_{m \in \mathcal{S} \cup \mathcal{T}} \sum_{k=1}^{K} \sum_{c=1}^{C} \mathcal{L}_{img}(\boldsymbol{y}_m(c), \hat{\boldsymbol{y}}_m^k(c)),$$

where $\mathcal{L}_{img}$ is the cross-entropy loss trained with soft labels. The image-level labels of the target images are produced by the multinomial logistic regression model trained with source images. The reconstruction loss involves both the source and target domains:

$$\mathcal{L}_r = \sum_{m \in \mathcal{S} \cup \mathcal{T}} d_m.$$

To sum up, with the above sub-objectives, our final loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r,$$

where λ_c and λ_r are trade-off parameters that weigh the importance of the corresponding terms.

5. Experiments

5.1. Experimental setup

We follow the common protocol of previous works [34, 5, 1] and conduct experiments on the standard benchmark settings, including both "GTAV to Cityscapes" and "SYNTHIA to Cityscapes". Specifically, we take GTAV or SYNTHIA datasets with pixel-level annotations as the source domain and Cityscapes dataset without any annotations as the target domain.

Cityscapes is a dataset that contains real-world photos from the urban traffic scenes of 50 different cities in Germany [9]. Its official data split includes 2,975 training images and 500 validation ones, each of which has a resolution of 2048×1024. There exist 19 possible semantic labels for each pixel. Following the prior works [1, 34], we evaluate the performance on the validation set of Cityscapes.

GTAV is a dataset that contains synthetic images rendered from the game engine of Grand Theft Auto V [27]. It includes 24,966 images with a resolution of 1914×1052. The pixel-level labels, which are generated automatically by computer graphics techniques, are fully compatible with Cityscapes. In experiments, 19 common categories between GTAV and Cityscapes are of interest.

SYNTHIA is another synthetic dataset that contains 9,400 photo-realistic images with a resolution of 1280×960 rendered from virtual scenes [28]. Like GTAV, the pixel-level labels are automatically produced and compatible with Cityscapes. Following the previous works [34, 5], we use the SYNTHIA-RAND-CITYSCAPES subset in the experiments and evaluate performance on 16 common categories between SYNTHIA and Cityscapes.

5.2. Implementation details

We conduct experiments employing deeplab-v2 with ResNet-101 [2] and FCN8s with VGG-16 [21], both of which are pre-trained on ImageNet, as the base networks. Similar to [34], the images of GTAV and Cityscapes are resized to the resolution of 1280×640 and 1024×512, respectively. The resolution of SYNTHIA images remains unchanged. To validate the robustness of our method, we adopt the same hyper-parameter setting in both "GTAV to Cityscapes" and "SYNTHIA to Cityscapes". The model is trained by the Stochastic Gradient Decent (SGD) optimizer with the initial learning rate of 2.5×10^{-5} and the momentum of 0.9. The maximum iteration number is set to 200,000 and the batch size is set to 4. The learning rate decreases according to the polynomial decay policy with power of 0.9. We use 3 region expansion units in our model, with the hyper-parameters r^k set to $\{0.5, 1, 1.5\}$. The values of λ_r and λ_c are set to 0.5.

The implementation of the logistic regression model follows [41]. In the regression head, there is only one layer. SGD optimizer with the initial learning rate of 1.0×10^{-4} and the momentum of 0.9 is used. The learning rate scheduler follows exponential decay policy with the decay rate of 0.99 and the minimum learning rate of 1.0×10^{-6}.

5.3. Performance comparison

Our proposed method is compared with the existing state-of-the-art methods, including [23, 35, 7, 37, 1, 24, 3, 36, 14, 47, 6, 34, 8, 13, 42, 39]. Of these, the works [23, 35, 42, 6, 8, 39] mainly focus on leaning domain-invariant representation of instances; the works [7, 14] directly translate the source images into target-style images with GANs; the works [37, 5, 1, 24, 34] exploit the structural consistency between domains through conducting domain-adversarial training at the output space of semantic segmentation networks; the works [3, 36, 47] are inspired by the semi-supervised learning methods such as self-training or entropy minimization; and Saleh et al. [13] stands out from the other works by domain flow generation. We adopt the mean Intersection-Over-Union (mIoU) value as the metric of evaluation.

GTAV to Cityscapes. We display the comparisons of unsupervised domain adaptation from GTAV to Cityscapes in Table 1. The following observations can be drawn. First, domain adaptation approaches can outperform the "source only" models that are trained with only the source images by a large performance gain. This observation validates the effectiveness of using domain adaptation techniques to promote the adaptation of semantic segmentation networks.

Table 1: Comparison on "GTAV to Cityscapes" in terms of per-class IoUs and mIoU (%).

Method	Base Model	road	sdwk	bldng	wall	fence	pole	light	sign	vgttn	trm	sky	person	rider	car	truck	bus	train	mcycl	bcycl	mIoU
Source only [39]	VGG-16	72.5	25.1	71.2	6.6	13.4	12.3	11.0	4.7	76.1	16.4	67.7	43.1	8.0	70.4	11.3	4.8	0.0	13.9	0.4	27.8
Source only(ours)		66.5	23.3	68.2	17.1	12.1	14.5	16.0	4.0	79.6	16.7	64.2	40.3	2.1	70.8	20.5	16.8	2.0	8.9	0.0	28.6
DPR [35]		87.3	35.7	79.5	**32.0**	14.5	21.5	24.8	13.7	80.4	**32.0**	70.5	50.5	16.9	81.0	20.8	28.1	4.1	15.5	4.1	37.5
ROAD [6]		85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
SIBAN [23]		83.4	13.0	77.8	20.4	17.5	24.6	22.8	9.6	81.3	29.6	77.3	42.7	10.9	76.0	22.8	17.9	5.7	14.2	2.0	34.2
DCAN [39]		82.3	26.7	77.4	23.7	20.5	20.4	**30.3**	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.3	17.0	6.7	36.2
CrDoCo [7]		89.1	33.2	80.1	26.9	25.0	18.3	23.4	12.8	77.0	29.1	72.4	**55.1**	20.2	79.9	22.3	19.5	1.0	20.1	18.7	38.1
CyCADA [14]		85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
AdaptSeg [34]		87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0
CLAN [24]		88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	**29.9**	0.0	10.7	0.0	36.6
AdvEnt [36]		86.9	28.7	78.7	28.5	**25.2**	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	**81.5**	26.0	17.2	18.9	11.7	1.6	36.1
CBST [47]		**90.4**	**50.8**	72.0	18.3	9.5	**27.2**	28.6	14.1	**82.4**	25.1	70.8	42.6	14.5	76.9	5.9	12.5	1.2	14.0	**28.6**	36.1
PIT(ours)		86.2	35.0	**82.1**	31.1	22.1	23.2	29.4	**28.5**	79.3	31.8	**81.9**	52.1	**23.2**	80.4	**29.5**	26.9	**30.7**	**20.5**	1.2	**41.8**
Source only [35]	ResNet-101	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Source only(ours)		75.2	20.2	77.7	22.6	20.9	25.7	27.8	18.3	80.1	9.8	73.1	56.0	23.0	65.4	27.0	6.8	2.7	21.8	34.3	36.2
DPR [35]		**92.3**	**51.9**	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	**46.3**	2.2	29.5	32.3	46.5
SIBAN [23]		88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	**40.0**	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
FCAN [42]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.6
AdaptSeg [34]		86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CLAN [24]		87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
DISE [1]		91.5	47.5	**82.5**	**31.3**	25.6	33.0	33.7	25.8	82.7	28.8	**82.7**	62.4	30.8	**85.2**	27.7	34.5	6.4	25.2	24.4	45.4
AdvEnt [36]		89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
MSL [3]		89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	**85.3**	39.4	78.2	63.0	22.9	84.6	36.4	43.0	5.5	**34.7**	33.5	46.4
DLOW [13]		87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5	42.3
PIT(ours)		87.5	43.4	78.8	31.2	**30.2**	**36.3**	**39.9**	**42.0**	79.2	37.1	79.3	**65.4**	**37.5**	83.2	**46.0**	45.6	**25.7**	23.5	**49.9**	**50.6**

Moreover, our method can achieve significantly better results than that of the compared state-of-the-art works. In particular, our method is good at predicting foreground objects, such as "light", "person", "motorcycle", "sign", "rider", "truck", "train", etc. It is because that the image-level predictions in corse-grained component can explicitly encourage the region activation for these foreground classes in the class activation map and thus improve the performance.

In Fig. 2, we further display the qualitative segmentation results. We can see that the segmentation masks obtained by our method can reflect a good understanding of the road traffic scenes. Both the dominant background classes (e.g. "road", "sidewalk" and "building") and foreground objects (e.g. "light", "person" and "car") can be well predicted.

SYNTHIA to Cityscapes. Comparisons are also conducted on the "SYNTHIA to Cityscapes" setting. From Table 2, it is clear that our method can outperform the compared methods with a large performance gain. We can draw the similar discussions as in "GTAV to Cityscapes". Our approach is also potential to be applied to other domain adaptation scenarios like semantic segmentation cross weather conditions, lighting conditions or cities in autonomous driving, due to the relatively constant appearance of each category.

5.4. Analysis

We further report the ablation study results to demonstrate the contributions of each design in our proposed method in Table 3. The first row displays the performance of the full PIT model. From the next two rows, it is clear that both the two losses, $\mathcal{L}_c$ and $\mathcal{L}_r$, are essential to obtain a good segmentation model. Specifically, for the second row, when the coarse-grained component does not learn from the image-level categories, the reconstruction loss cannot capture correct interactions for transfer. For the third row, when the fined-grained and coarse-grained components do not interact with each other, the model will fail to exploit the underlying domain-invariant correlations between them. Furthermore, we evaluate the performance decoupling the source images from the reconstruction loss, which is displayed in the fourth row. It is clear that this will cause a significant performance drop, which demonstrates that the interactions between the two different granularity levels are mainly captured by the source domain, and then transferred to the target domain. This result confirms our assumption that the interactive relation between the image-level and pixel-level category information is the pivot information shared by different domains. In the next three rows, we evaluate the performance using one expansion unit to model the interactive relation. We can see that only one region expansion unit is insufficient to exploit the pivot knowledge for transfer. We draw two reasons for this. First, with one region expansion unit, the reconstruction loss can only model a very simple interactive relation, i.e. a scaling relation, between the class activation maps of the two components. Second, for urban traffic scenes, since the region size varies a lot across different categories, one single region expansion unit can only capture some aspect of the pixel-level

Table 2: Comparison on "SYNTHIA to Cityscapes" in terms of per-class IoUs and mIoU (%). The mIoU* column denotes the mean IoU over 13 categories excluding those marked by *.

Method	Base Model	road	sdwk	bldng	wall*	fence*	pole*	light	sign	vgttn	sky	person	rider	car	bus	mcycl	bcycl	mIoU	mIoU*
Source only [39]	VGG-16	10.8	11.4	66.6	1.6	0.1	16.9	5.5	14.1	74.2	76.2	46.0	11.5	45.4	15.1	6.0	13.4	25.9	30.4
Source only(ours)		8.6	9.6	76.3	1.3	0.5	5.5	6.7	3.9	74.3	71.3	39.9	7.3	70.0	17.4	3.5	6.2	25.1	30.3
DPR [35]		72.6	29.5	77.2	3.5	0.4	21.0	1.4	7.9	73.3	79.0	45.7	14.5	69.4	19.6	7.4	16.5	33.7	39.6
SIBAN [23]		70.1	25.7	**80.9**	-	-	-	3.8	7.2	72.3	80.5	43.3	5.0	73.3	16.0	1.7	3.6	-	37.2
DCAN [39]		79.9	30.4	70.8	1.6	**0.6**	22.3	6.7	**23.0**	76.9	73.9	41.9	16.7	61.7	11.5	10.3	**38.6**	35.4	41.6
CWDA [8]		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
AdaptSeg [34]		78.9	29.2	75.5	-	-	-	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	-	37.6
CLAN [24]		80.4	**30.7**	74.7	-	-	-	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	-	39.3
AdvEnt [36]		67.9	29.4	71.9	6.3	0.3	19.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	31.4	36.6
CBST [47]		69.6	28.7	69.5	**12.1**	0.1	**25.4**	11.9	13.6	**82.0**	**81.9**	**49.1**	14.5	66.0	6.6	3.7	32.4	35.4	36.1
PIT(ours)		**81.7**	26.9	78.4	6.3	0.2	19.8	**13.4**	17.4	76.7	74.1	47.5	**22.4**	**76.0**	**21.7**	**19.6**	27.7	**38.1**	**44.9**
Source only [35]	ResNet-101	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
Source only(ours)		57.4	21.5	74.6	3.2	0.7	4.6	7.8	9.7	72.6	80.0	53.7	21.8	38.5	22.1	11.8	24.8	31.5	38.2
DPR [35]		82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	**84.6**	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
SIBAN [23]		82.5	24.0	79.4	-	-	-	16.5	12.7	79.2	82.8	58.3	18.0	79.3	25.3	17.6	25.9	-	46.3
AdaptSeg [34]		84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
CLAN [24]		81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
DISE [1]		**91.7**	**53.5**	77.1	2.5	0.2	**27.1**	6.2	7.6	78.4	81.2	55.8	19.2	**82.3**	30.3	17.1	34.3	-	41.5
DADA [37]		89.2	44.8	81.4	6.8	0.3	26.2	8.6	11.1	81.8	84.0	54.7	19.3	79.7	**40.7**	14.0	**38.8**	42.6	49.8
AdvEnt [36]		85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
MSL [3]		82.9	40.7	80.3	**10.2**	**0.8**	25.8	12.8	18.2	**82.5**	82.2	53.1	18.0	79.0	31.4	10.4	35.6	41.4	48.2
PIT(ours)		83.1	27.6	**81.5**	8.9	0.3	21.8	**26.4**	**33.8**	76.4	78.8	**64.2**	**27.6**	79.6	31.2	**31.0**	31.3	**44.0**	**51.8**

Table 3: Ablation study of the proposed PIT in terms of mIOU (%). The notation $-\mathcal{L}_r(\mathcal{S})$ indicates decoupling the source data from the reconstruction loss.

Model design	GTAV		SYNTHIA	
	VGG-16	Res-101	VGG-16	Res-101
Full model	**41.8**	**50.6**	**38.1**	**44.0**
-$\mathcal{L}_c(\mathcal{S}+\mathcal{T})$	35.9	40.8	34.8	36.2
-$\mathcal{L}_r(\mathcal{S}+\mathcal{T})$	34.2	38.0	32.1	37.3
-$\mathcal{L}_r(\mathcal{S})$	36.4	38.5	35.2	39.9
Unit 1 ($r=0.5$)	38.2	44.9	35.3	40.1
Unit 2 ($r=1$)	39.2	45.2	34.8	41.1
Unit 3 ($r=1.5$)	37.5	45.7	36.1	39.6

Table 4: Comparisons with different number of region expansion units in terms of mIOU (%).

Number (r^k)	GTAV		SYNTHIA	
	VGG-16	Res-101	VGG-16	Res-101
1 ({0.5})	38.2	44.9	35.3	40.1
2 ({0.5,1})	40.2	48.2	37.3	42.5
3 ({0.5,1.0,1.5})	**41.8**	50.6	**38.1**	**44.0**
4 ({0.5,1.0,1.5,2.0})	41.3	**50.7**	37.0	43.5

ground truth. Hence, the coarse-grained component cannot learn enough information for transfer. We display the class activation map of each region expansion unit in Fig. 3. It is clear that the class activation map of each region expansion unit can reveal some aspect of the pixel-level ground truth, but none of them is absolutely accurate.

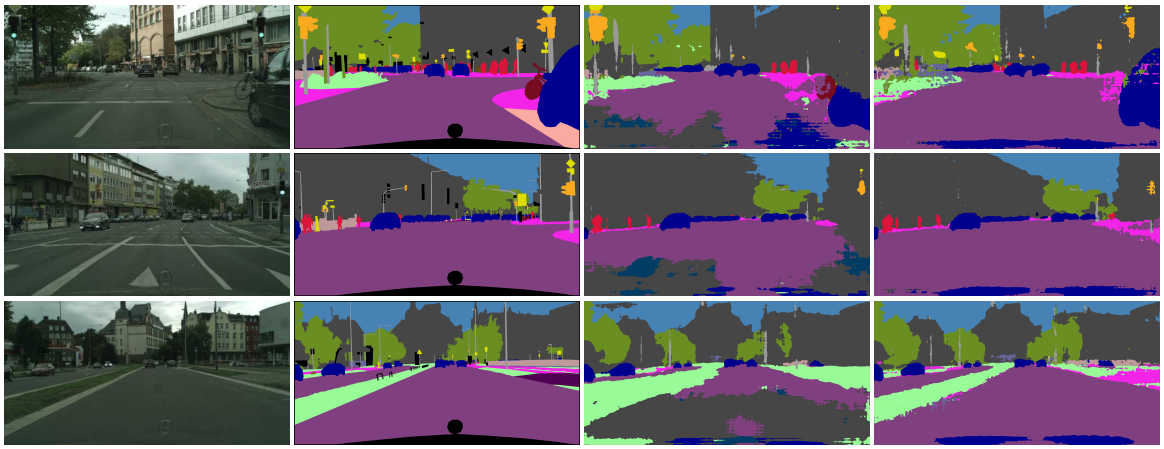
Moreover, we conduct comparisons with different number of region expansion units as shown in Table 4. We can see that the performance is improved gradually when we increase the number of region expansion units from 1 to 3. The performance begins to drop when we use 4 region expansion units. We ascribe it to the reason that the interactions between image-level and pixel-level information have already been captured by using 3 region expansion units. It will introduce more parameters and cause the problem of overfitting if more expansion units are used.

Table 5: Comparisons of different domain adaptation settings in terms of mIOU (%). UDA denotes the unsupervised domain adaptation setting and WDA denotes the weakly-supervised domain adaptation setting.

Setting (method)	GTAV		SYNTHIA	
	VGG-16	Res-101	VGG-16	Res-101
UDA (ours)	41.8	50.6	38.1	44.0
WDA (ours)	**43.8**	**52.9**	**42.4**	**47.3**
WDA (ODC [38])	37.4	-	35.7	-

5.5. Weakly-supervised domain adaptation

The accuracy of the image-level categories of the target domain is essential to obtain good performance. Obviously, our method is also suitable for the weakly-supervised domain adaptation setting, in which the image-level labels of the target domain are available during training. The image-level labels reveal whether a specific category exists in a target image. We report the comparisons of different domain adaptation settings in Table 5. We can see that our method obtains even better results given the image-level ground truth of target images. However, the performance of the unsupervised setting is generally close to that of the weakly-supervised setting. This reflects that the multinomial logistic regression model can predict the image-level



(a) target image (b) ground truth (c) source only (d) PIT

Figure 2: Qualitative results on "GTAV to Cityscapes", (a) target images, (b) ground truth, (c) segmentation results of the "source only" baseline, (d) segmentation results of the proposed PIT method.

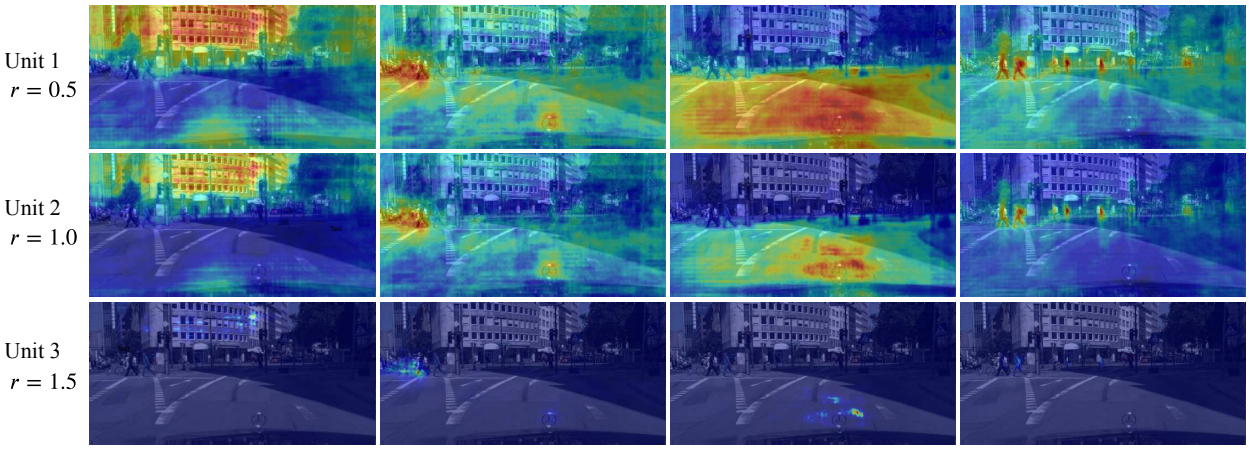


Figure 3: The class activation map of each region expansion unit.

categories of target images well. Additionally, compared to Objects Domain Classifier (ODC) that utilizes box-level label from the target domain as weak supervision [38], our method obtains better results, even though the box-level labels convey much more information than image-level weak supervision.

6. Conclusion

In this paper, we propose a new domain adaptation approach by constructing pivot information that is common knowledge shared across domains to promote the adaptation of semantic segmentation neural networks. To this end, we design a multi-level region expansion mechanism to model the domain-invariant interaction between the image-level categories and the pixel-level semantic information. Unlike the prior works that focus on learning domain-invariant representations of instances by using domain adversarial training, our method constructs the transferrable factors across different domains explicitly. Additionally, our method does not rely on the adversarial training which is often hard to optimize. We conduct extensive experiments on the adaptation from both GTAV and SYNTHIA to Cityscapes. The experimental results clearly demonstrate that our method obtains better results than the existing state-of-the-art works.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No.11829101 and 11931014) and the Fundamental Research Funds for the Central Universities of China (No.JBK1806002). G. Lin's participation was supported by the National Research Foundation Singapore under its AI Singapore Programme [AISG-RP-2018-003]. F. Lv devotes this work to his father for the thanks to his parenting and company.

References

[1] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: adapting structural information across domains for boosting semantic segmentation. In *CVPR*, pages 1900–1909, 2019.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018.

[3] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, pages 2090–2099, 2019.

[4] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *CVPR*, pages 7976–7985, 2018.

[5] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: a geometrically guided input-output adaptation approach. In *CVPR*, pages 1841–1850, 2019.

[6] Yuhua Chen, Wen Li, and Luc Van Gool. Road: reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, pages 7892–7901, 2018.

[7] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: pixel-level domain transfer with cross-domain consistency. In *CVPR*, pages 1791–1800, 2019.

[8] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: cross city adaptation of road scene segmenters. In *ICCV*, pages 2011–2020, 2017.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[10] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: separated semantic feature based domain adaptation network for semantic segmentation. In *ICCV*, pages 982–991, 2019.

[11] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[13] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: domain flow for adaptation and generalization. In *CVPR*, pages 2477–2486, 2019.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[15] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[16] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: three principles for weakly-supervised image segmentation. In *ECCV*, pages 695–711, 2016.

[17] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: learning discriminative features for unsupervised domain adaptation. In *ICCV*, pages 91–100, 2019.

[18] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019.

[19] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017.

[20] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[23] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. *ICCV*, 2019.

[24] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation. In *CVPR*, pages 2507–2516, 2019.

[25] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *CVPR*, pages 4500–4509, 2018.

[26] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721, 2015.

[27] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: ground truth from computer games. In *ECCV*, pages 102–118, 2016.

[28] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, pages 3234–3243, 2016.

[29] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, pages 2988–2997, 2017.

[30] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018.

[31] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, pages 7374–7383, 2019.

[32] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *arXiv preprint arXiv:1711.06969*, 2017.

[33] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[34] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018.

[35] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *ICCV*, 2019.

[36] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019.

[37] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Dada: depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019.

[38] Qi Wang, Junyu Gao, and Xuelong Li. Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Trans. Image Processing*, 28(9):4376–4386, 2019.

[39] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *ECCV*, pages 518–534, 2018.

[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[41] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[42] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, pages 6810–6818, 2018.

[43] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, pages 5031–5040, 2019.

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[45] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015.

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[47] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018.