ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes

Yuhua Chen[1]   Wen Li[1]   Luc Van Gool[1,2]

[1]Computer Vision Laboratory, ETH Zurich   [2]VISICS, ESAT/PSI, KU Leuven

{yuhua.chen,liwen,vangool}@vision.ee.ethz.ch

Abstract

Exploiting synthetic data to learn deep models has attracted increasing attention in recent years. However, the intrinsic domain difference between synthetic and real images usually causes a significant performance drop when applying the learned model to real world scenarios. This is mainly due to two reasons: 1) the model overfits to synthetic images, making the convolutional filters incompetent to extract informative representation for real images; 2) there is a distribution difference between synthetic and real data, which is also known as the domain adaptation problem. To this end, we propose a new reality oriented adaptation approach for urban scene semantic segmentation by learning from synthetic data. First, we propose a target guided distillation approach to learn the real image style, which is achieved by training the segmentation model to imitate a pretrained real style model using real images. Second, we further take advantage of the intrinsic spatial structure presented in urban scene images, and propose a spatial-aware adaptation scheme to effectively align the distribution of two domains. These two modules can be readily integrated with existing state-of-the-art semantic segmentation networks to improve their generalizability when adapting from synthetic to real urban scenes. We evaluate the proposed method on Cityscapes dataset by adapting from GTAV and SYNTHIA datasets, where the results demonstrate the effectiveness of our method.

1. Introduction

With the exciting vision of autonomous driving, semantic segmentation of urban scenes, as a key module, has gained increasing attention from both academia and industry. However, collecting and labeling training data for semantic segmentation task is a laborious and expensive process, as it requires per-pixel annotation. This issue becomes even more severe with the surge of deep learning techniques, which usually require a large amount of training data. Therefore, it becomes much desired to exploit low cost ways to acquire data for semantic segmentation.

One way that becomes recently prevalent is to collect photo-realistic synthetic data from video games, where pixel-level annotation can be automatically generated at a much lower cost. For example, Richter *etal.*[32] constructed a large scale synthetic urban scene dataset for semantic segmentation from the GTAV game. While the cost of acquiring training data and annotation is largely reduced, synthetic data still suffers from a considerable domain difference from the real data, which usually leads to a significant performance drop when applying the segmentation model to real world urban scenes [17, 49].

The main reasons are two-fold. First, from the perspective of representation, since the model is trained on synthetic images, the convolutional filters tend to overfit to synthetic style images, making them incompetent to extract informative features for real images. Second, from the distribution perspective, synthetic and real data suffers a considerable distribution mismatch, which makes the model biased to synthetic domain.

To overcome such problems, we propose **R**eality **O**riented **AD**aptation Networks(ROAD-Net) for semantic segmentation of urban scenes by learning from synthetic data. We address the above two issues respectively by a target guided distillation module for real style orientation, and a spatial-aware adaptation module for real distribution orientation, which are described respectively as follows,

- **real style orientation:** To prevent the segmentation model from overfitting to synthetic images, we propose to use the target real images to imitate a pretrained real style model. This can be achieved using the model distillation strategy by enforcing the output from segmentation model similar with the output of a pretrained model. On one hand, this encourages convolutional filters to fit to the real images through the distillation task. On the other hand, it also enforces the segmentation network to preserve good discriminative for real images by approaching the semantic output from the pretrained model. We refer to it as *target guided distillation.*

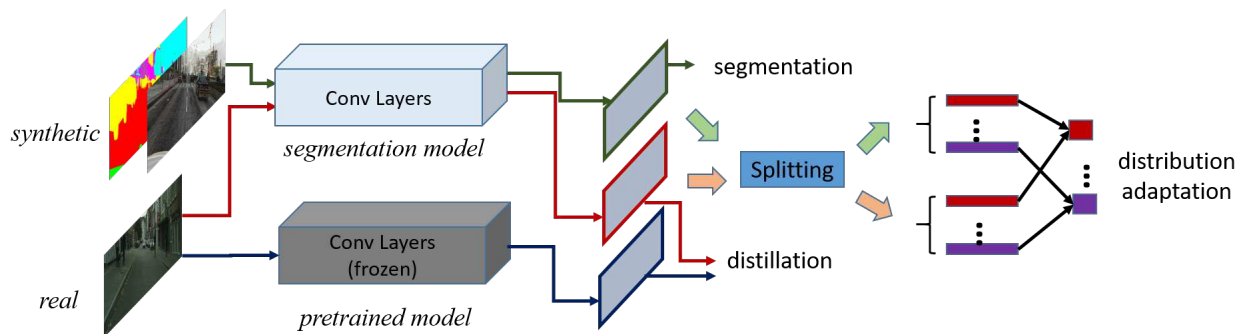- **real distribution orientation:** To deal with the dis-

Figure 1. Illustration of our Reality Oriented Adaptation networks(ROAD-Net) for semantic segmentation of urban scenes. Our network is built upon conventional semantic segmentation networks, and incorporates a *target guided distillation* module for real style orientation, and a *spatial-aware adaptation* module for real distribution orientation.

tribution mismatch for semantic segmentation, several recent works [17, 49] applied domain adaptation methods on pixel-level features. However, this is generally a challenging task partially due to the large visual variance in urban scene. For example, the objects in the central region are usually much smaller, compared to objects in the outer region. Aligning distributions directly under such large variance presents significant difficulty. We therefore aim to ease the domain adaptation task by exploiting the intrinsic geometry information presented in urban scene. We propose a *spatial-aware adaptation* method to effectively align the two domains while considering the difference in spatial distribution. In particular, we divide the scene image into different spatial regions, and align the source and target domain samples from the same spatial region respectively. In this way, we align features from two domains with similar spatial properties.

The two modules above can be easily integrated with existing state-of-the-art semantic segmentation networks to boost their generalizability when adapting from synthetic to real urban scenes. We conduct extensive experiments by using the GTAV dataset and the Cityscapes datset. Our proposed method achieves a new state-of-the-art of 39.4% mean IoU. To further validate the effectiveness, we additionally evaluate our methods using SYNTHIA dataset and the Cityscapes dataset, in which the proposed method outperforms other competing methods as well.

2. Related Works

Semantic Segmentation: Semantic segmentation is a highly active field since decades ago with large amount of methods proposed, we briefly review some of the works with a focus on CNN-based methods. Traditional works in semantic segmentation [37, 42, 48] are typically based on manually designed image features. With the recent surge of deep learning [19], learned representation demonstrated its

power in many computer vision tasks. Pioneered by [25], the power of CNN has been transferred to semantic segmentation and we have witnessed a rapid boost in semantic segmentation performance. Long *etal.*[25] formulates semantic segmentation as a per-pixel classification problem. Following the line of FCN-based semantic segmentation. Dilated convolution is proposed by [47] to enlarge the receptive field. DeepLab [1] incorporates Conditional random filed(CRF) with CNN to reason about spatial relationship. Recently, Zhao *etal.*[50] proposed to use *Pyramid Pooling Module* to encode the global and local context, which achieved state-of-the-arts results on multiple datasets.

Domain Adaptation: A basic assumption in conventional machine learning is that the training and test data are sampled independently from an identical distribution, or *i.i.d* assumption in short. However, this does not always hold in real world scenarios, which often leads to a significant performance drop on the test data when applying the trained model. Domain adaptation aims to alleviate the impact of such distribution mismatch such that the generalization ability of the learned model can be improved on the target domain. In computer vision, domain adaptation has been widely studied as an image classification problem in computer vision [20, 12, 10, 7, 40, 26, 8, 9, 36, 30, 29, 23]. Conventional methods include asymmetric metric learning [20], subspace interpolation [12], geodesic flow kernel [10], subspace alignment [7], covariance matrix alignment [40], *etc*. Recent works aim to improve the domain adaptability of deep neural networks, including [26, 8, 9, 36, 30, 29, 22, 14, 27, 28]. Different from those works, our work aims to solve the semantic segmentation task, which is more challenging due to the large variance in pixel-level features.

Domain Adaptation for Semantic Segmentation: So far most works in domain adaptation focus on the task of image classification. Not until recently has the community pay attention to domain shift problem in semantic segmentation. This line of research is pioneered by [17], where

they use adversarial training to align the feature from both domain. Another approach is done by [49], where they use a curriculum learning style approach to solve the domain shift. There are also some concurrent works [35, 34, 4, 43] concerning similar problem in different ways.

Learning Using Synthetic Data: There are also a few works proposed to learn from synthetic data [41, 46, 33, 44, 31, 38]. In [41], generic object detectors are trained from synthetic images, while in [44] virtual images are used to improve pedestrian detections in real environment. More recently, [3] extends Faster R-CNN to learn a domain-invariant detector using synthetic data or data from another domain.

Other Related Works : From the methodology, our work is inspired by the model distillation [16], which was proposed for network compression. The following up works employed a similar strategy to perform knowledge transfer [24, 45], and cross-modality supervision transfer [13]. Our work, on the other hand, using the distillation strategy for learning the real style convolutional filters for semantic segmentation. Our work is also inspired by the previous works on scene understanding [21, 18], in which they also exploited the intrinsic spatial layout prior of scene images for different tasks in various ways. We share a similar philosophy with those work, but aim to solve the distribution alignment between two domains for the urban scene semantic segmentation, which results in a totally different methodology.

3. Reality Oriented Adaptation Networks

In this section, we present our Reality Oriented Adaptation networks(ROAD-Net) for urban scene semantic segmentation. Two new modules are proposed in our ROAD-Net model, the *target guided distillation* and *spatial-aware adaptation*, which will be introduced respectively in this section.

3.1. Target Guided Distillation

When learning a model from synthetic urban scenes, one major issue is that synthetic data usually exhibits a clear visual difference with real images. As a result, the learned model is biased towards the synthetic data, which leads to an unsatisfactory segmentation performance on real urban scene images.

From the perspective of representation learning, this is largely because that the convolutional filters in the semantic segmentation models overfit to the synthetic image style. Consequently, when taking a real image as input, the feature representation generated using the learned model might not be sufficiently informative for semantic segmentation, since the convolutional filters tend to extract information that look discriminative in synthetic style only. To cope with this issue, we therefore propose to use the target real images to
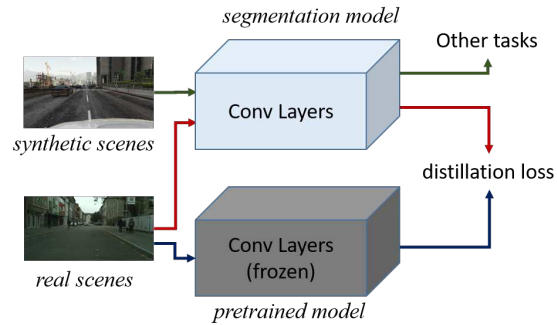

Figure 2. Illustration of target guided distillation. By feeding the target real images into both the segmentation model, and a pretrained model, the segmentation model is encourage to imitate the pretrained model to learn real style convolutional filters.

guide the segmentation model to learn robust convolutional filters for the real urban scenes.

Our method is motivated by the common practice in computer vision community of initializing network weights using a network [15, 39] pretrained on large-scale image dataset such as ImageNet [6]. Such strategy has also been widely exploited for learning semantic segmentation models [1, 47, 50]. Recall that the original ImageNet model is pretrained on real images, so we propose to employ a distillation loss to guide the semantic segmentation models to behave like the pretrained real style model.

We illustrate the target guided distillation process in Figure 2. The source synthetic urban scene images and corresponding annotation are used to learn the semantic segmentation model. At the same time, the unlabeled real urban scene images from target domain are used to imitate their corresponding feature map output from the pretrained ImageNet model. The pretrained model has the same structure as the backbone network used in the semantic segmentation model, and is frozen during this process.

Formally, for a real image, let us denote $\mathbf{x}_{i,j}$ as the activation at the position (i, j) of the feature map from the semantic segmentation model, and also denote $\mathbf{z}_{i,j}$ as an activation at the same location of the feature map from the pretrained model, then the loss for target guided distillation can be written as,

$$\mathcal{L}_{dist} = \frac{1}{N} \sum_{i,j} \|\mathbf{x}_{i,j} - \mathbf{z}_{i,j}\|_2 \qquad (1)$$

where N is the number of activations at the feature map, and $\| \cdot \|_2$ is the Euclidean distance.

Discussion: There also exists other alternative approaches to prevent convolutional filters from overfitting to synthetic data. For instance, considering the semantic segmentation model is initialized with a pretrained ImageNet model which is learned from the real images, a possible way

is to freeze a few convolutional layers (*i.e.*, to set the learning rates for those layers to zero), such that convolutional filters will not be corrupted by synthetic data. As shown in our experiments(*c.f.* Section 4.3), this does work to some extent. However, it is still inferior to our target guided distillation approach. The main reason is that the ImageNet model is trained for image classification, and may not be optimal to the semantic segmentation of urban scenes. Freezing a few layers may help to prevent the model from overfitting to synthetic data, but also limits its capacity for the semantic segmentation task. As a comparison, in target guided distillation, all weights are allowed to be tuned, and convolutional filters are guided in a soft manner to imitate the ImageNet model when being trained for the segmentation task.

Another alternative approach is to use source data to perform the distillation task, which is also known as learning without forgetting [24]. Though it is able to imitate the ImageNet pretrained model, we argue that it is not as effective as using the target real urban scene images, because the convolutional filters can be still corrupted due to taking solely synthetic images as input. Moreover, the deep neural networks often contain multiple layers, and the filters in higher layers are able to be trained to well imitate the ImageNet model even with corrupted low layers convolutional filters. We conduct an experimental comparison with all alternative baselines in Section 4.3.

3.2. Spatial-Aware Adaptation

Even with the target guided distillation, the feature presentation of two domains extracted by the semantic segmentation networks could still suffer from a distribution mismatch, due to the large domain difference between synthetic and real images. Thus, it is desirable to align the distributions of two domains with domain adaptation methods.

However, aligning pixel-level features between synthetic and real data for urban scene images is non-trivial. Conventional domain adaptation approaches are usually proposed for the image classification task. While similar methodology can be applied by taking each pixel-level feature as a training sample, it is still challenging to fully reduce the distribution mismatch, since the pixels vary significantly in appearance [17] and scale [2].

To address this issue, we hence propose to leverage the intrinsic spatial structure in urban scene images to improve the domain distribution alignment between synthetic and real data. Our motivations are two-fold. Firstly, while the pixel-level features vary a lot, such variance also exhibits a nice spatial pattern. For example, in an urban scene image, the objects in the central region are usually small, while objects in the outer region are relatively larger. Second, the semantic categories also roughly follow a spatial layout. Usually *road* appears in the bottom part of an image,
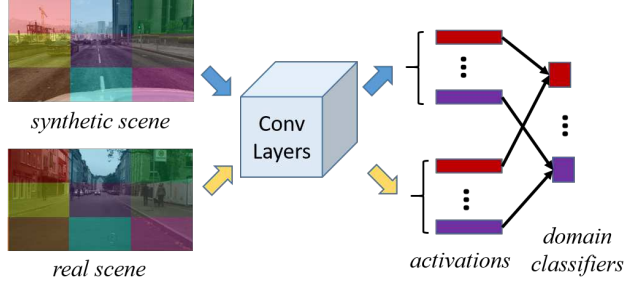

Figure 3. Illustration of spatial-aware adaptation. The source and target pixel-level features in the same spatial region are aligned(*i.e.*, the regions marked with the same color), which eases the domain distribution adaptation task.

and *sky* appears on the top part of an image. Therefore, it is beneficial to align pixel-level features with similar sizes and semantics based on their spatial location.

In particular, we divide each scene image into different spatial regions, and domain alignment is performed on the pixel-level features from the same spatial region. We illustrate this process in Figure 3. For each region, any traditional domain distribution alignment can be deployed. Formally, suppose there are $m = 1, \ldots, M$ regions. We denote $\mathcal{R}_m$ as the set of indices contains in the m-th region, *i.e.*, $(u, v) \in \mathcal{R}_m$ means that a pixel-level activation $\mathbf{x}_{u,v}$ locates at this region, and vice versa. Let us denote $\mathcal{X}_m^s = \{\mathbf{x}_{u,v}^s | (u, v) \in \mathcal{R}_m\}$ and $\mathcal{X}_m^t = \{\mathbf{x}_{u,v}^t | (u, v) \in \mathcal{R}_m\}$ as all pixel-level activations locate at the m-th region for source and target domains respectively, then the loss for spatial-aware adaptation can be written,

$$\mathcal{L}_{spt} = \sum_{m=1}^{M} \mathcal{L}_{da}(\mathcal{X}_m^s, \mathcal{X}_m^t) \qquad (2)$$

where $\mathcal{L}_{da}$ is a domain adaptation loss which measures the domain difference between two sets of samples. We present an example of $\mathcal{L}_{da}$ below.

Domain Adversarial Training: Aligning two distributions has been widely studied in the literature. In this work, we deploy an $\mathcal{H}$-divergence based loss used in the DANN model [8]. Specifically, let us denote $h : \mathbf{x} \to \{0, 1\}$ as a domain classifier, which is used to predict which domain an input pixel-level feature $\mathbf{x}$ comes from, where 0 denotes the source domain, and 1 denotes the target domain. Given a set of training data $\mathcal{X} = \mathcal{X}^s \cup \mathcal{X}^t$, the loss for training the domain classifier h can be written as:

$$\mathcal{L}_{\mathcal{H}}(\mathcal{X}^s, \mathcal{X}^t) = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \ell(h(\mathbf{x}), d) \qquad (3)$$

where $|\mathcal{X}|$ is the number of samples in $\mathcal{X}$, $d \in \{0, 1\}$ is the domain label of $\mathbf{x}$, $\ell(\cdot, \cdot)$ is a conventional classification loss for which we use the cross-entropy loss.

Intuitively, training a domain classifier is to distinguish samples from two domains. To reduce the domain difference, we thus encourage the activation $\mathbf{x}$ to be domain-indistinguishable. Considering each $\mathbf{x}$ is generated from a base network, denoted by F, we thus need to optimize F such that the domain classification loss $\mathcal{L}_{\mathcal{H}}$ is maximized. By jointly learning the domain classifier h and the base network F, we arrive at the following maxmin problem,

$$\max_F \min_h \mathcal{L}_{\mathcal{H}}(\mathcal{X}^s, \mathcal{X}^t) \qquad (4)$$

The above maxmin problem can be optimized in an adversarial training manner. Hoffman *et al.* [17] implemented it by switching domain label similarly as in the Generative Adversarial Networks (GAN) [11]. We follow [8] to insert a gradient reverse layer between F and h. Particularly, in back-propagation, the sign of gradients will be flipped when being passed through the gradient reverse layer, hence the classifier h is minimized while the network F maximized by directly using the conventional optimization methods like stochastic gradient descent (SGD).

3.3. Network Overview

The aforementioned two modules can be integrated with a conventional semantic segmentation network, such as DilatedNet [47], DeepLab [1], PSPNet [50] *etc*. In particular, let us denote $\mathcal{L}_{seg}$ as the segmentation loss, then our ROAD-Net model can be trained by minimizing a joint loss as follows,

$$\mathcal{L}_{ROAD} = \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{dist} + \lambda_2 \mathcal{L}_{spt}. \qquad (5)$$

where λ_1 and λ_2 are two trade-off parameters, which are set to 0.1 and 0.01 in our experiments. Note that the loss $\mathcal{L}_{spt}$ in the above equation corresponds to $\mathcal{L}_{\mathcal{H}}$ in (4). Although we have removed the max operator in (4) from the above loss, it can be automatically achieved with the reverse gradient layer.

We illustrate the pipeline of our proposed ROAD-net in Fig 1. During training, both synthetic and real images are fed into the network as input. The synthetic images and the annotation are used to train the segmentation task, while the real images are used to train the target guided distillation task. Both synthetic and real images are used for optimizing the spatial-aware adaptation loss. Note that, when training semantic segmentation model, due to the high resolution of input image, the input image is usually randomly cropped to fit the GPU memory. To perform the spatial-aware adaptation, we therefore build a spatial-aware splitting layer, in which we recover the location for each activation in the original image coordinate, and split them into different domain classifiers according to the region it comes from. During the test phase, the two newly added modules can be removed, and one can perform semantic segmentation as the same as for the conventional semantic segmentation models.

4. Experimental Results

In this section, we present our experimental results on semantic segmentation of real urban scenes by learning from synthetic data. Experimental analysis and comparison with state-of-the-arts are also provided.

4.1. Experiment Setup

We follow the classical unsupervised domain adaptation protocol where the supervision is assumed to be given in the source domain, and only unlabeled data is provided in the target domain. Following previous works [17, 49], the experimental validation are conducted on the GTAV dataset [32] and Cityscapes dataset [5]. We use GTAV dataset as our source domain, and we have access to the pixel-level annotation, and we use Cityscapes dataset as the target domain. Our goal is to learn a semantic segmentation model from synthetic data. We briefly introduce the datasets used in our experiment in below:

Cityscapes is a dataset focused on autonomous driving, which consists of $2,975$ images in training set, and 500 images for validation. The images have a fixed resolution of 2048×1024 pixels, and are captured by a video camera mounted in front of a car. 19 semantic categories are provided with pixel-level labels. In our experiment, we use the unlabeled images from the training set as the target domain to adapt our segmentation model, and the results are reported on the validation set.

GTAV is a dataset recently proposed for learning from synthetic data. It has $25,000$ photo-realistic images rendered by the gaming engine Grand Theft Auto V (GTAV). The resolution of images is around 2000×1000 pixels which is similar with Cityscapes, the semantic categories are also compatible between the two datasets.

In our experiments, we test the our proposed ROAD-Net with two semantic segmentation methods: DeepLab v2 [1] and PSPNet [50]. However, as discussed in the previous section, our method can be straightforwardly applied to other semantic segmentation methods as well.

The network is initialized using an ImageNet pre-trained weights, and the last convolutional layer is replaced with our new classification head to predict Cityscapes label. Similar with [1], a learning rate of 2.5×10^{-4} is used, and learning rate policy of *poly* is used in our experiment. Each batch contains 10 sampled patches of size 321×321, of which 5 patches are from source domain, and the other 5 patches from target domain. The cross-entropy loss is used for supervising semantic segmentation in the source domain as in [25, 1].

		dst	spt	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mean IoU
DeepLab	NonAdapt			29.8	16.0	56.6	9.2	17.3	13.5	13.6	9.8	74.9	6.7	54.3	41.9	2.9	45.0	3.3	13.1	1.3	6.8	0.0	21.9
	Ours	✓		82.0	29.5	73.8	20.0	12.7	8.5	14.6	9.0	73.2	27.6	63.2	39.3	5.9	71.7	23.1	17.6	**14.4**	4.0	0.0	31.1
			✓	85.3	**36.1**	77.5	24.1	18.8	21.4	22.5	**13.2**	71.3	20.9	54.5	40.6	6.9	76.7	**19.3**	16.4	7.0	7.1	0.0	32.6
		✓	✓	**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**
PSPNet	NonAdapt			14.6	17.0	40.4	21.1	14.7	18.3	18.0	5.6	80.8	16.0	68.4	49.1	4.4	78.5	**31.8**	23.9	1.5	24.4	0.0	27.8
	Ours	✓		73.2	32.0	71.0	25.2	**31.2**	15.7	14.4	**15.2**	82.5	**36.1**	73.1	52.1	17.1	78.1	26.5	28.0	3.1	24.7	4.4	36.8
			✓	75.9	34.2	**72.3**	28.0	21.7	**30.4**	25.2	**15.2**	79.4	33.5	**73.5**	52.3	**18.7**	76.2	31.0	25.2	3.5	23.7	2.1	37.8
		✓	✓	**76.3**	**36.1**	69.6	**28.6**	22.4	28.6	**29.3**	14.8	82.3	35.3	72.9	**54.4**	17.8	**78.9**	27.7	**30.3**	**4.0**	**24.9**	**12.6**	**39.4**

Table 1. The segmentation results on the Cityscapes dataset by using the GTAV dataset as the source domain. DeepLab and PSPNet are used as the base model, respectively. We report different variants of our proposed ROAD-Net, where "dst" and "spt" in the head row refers to the *target guided distillation* module and the *spatial-aware adaptation* module, respectively. The "NonAdapt" refers to the vanilla model which is trained using the GTAV data only. The best results for each base model are denoted in bold.

4.2. Experimental Results

For a comprehensive study on our method, we include two variants of our method, using only the *target guided distillation* module (referred to as "dst") and using only the *spatial-aware adaptation* module (referred to as "spt"). The vanilla base network without using any module is also included as a baseline (referred to as "NonAdapt"). Two different semantic segmentation networks are used, DeepLab V2 and the PSP Net, where the former one uses VGG-16 as the backbone network, and the latter one uses ResNet-101.

The results of all variants of our methods based on two different models are summarized in Talbe 1. Taking the results using DeepLab model as an example, the "NonAdapt" baseline gives 21.9% mean IoU, which is similar to those reported in [17, 49], where VGG-16 was also used as the backbone network. By using our proposed *target guided distillation* module and *spatial-aware adaptation* module respectively, we obtain a mean IoU of 31.1% and 32.6%, which improves the baseline by a large margin of 9.2% and 10.7%, respectively. This clearly demonstrates the effectiveness of the two modules for orienting the semantic segmentation model to learn reality. By combining two modules, our final ROAD-Net gains further improvements, achieving 35.9% in terms of mean IoU. Those observations can also be observed for the results using the PSP-Net. Moreover, using PSPNet, all methods are improved, and our ROAD-net finally achieves 39.4% mean IoU. Some randomly selected qualitative examples are shown in Fig. 5. We observe that the proposed method, especially the spatial aware module, improves the visual segmentation result notably, and produces predictions with much more reasonable spatial layout.

The proposed method does not achieve significant improvement on some categories, such as *pole* and *traffic sign*. A possible reason might be that those categories have very few pixels(*e.g.* in Cityscapes Val, only $1.8M$ pixels for *traffic light*, as a comparison, there are $345M$ pixels for *road*),
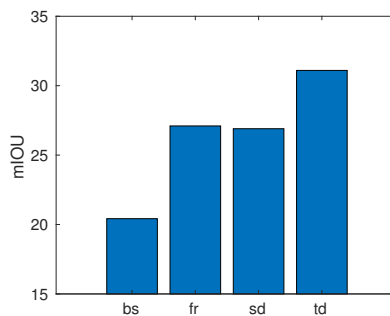

Figure 4. Study on alternative methods to target guided distillation: "bs" refers to the basline NonAdapt method, "fr" refers to the frozen method, "sd" refers to the source distillation method, and "td" refers to the target guided distillation.

which makes the results less stable compared to categories with more pixels.

4.3. Analysis on Real Style Orientation

To validate the effectiveness of our proposed target guided distillation module, we conduct experiments by comparing with the two alternative methods discussed in Section 3.1. In particular, the "**frozen**" method refers to freezing the first a few layers when training the segmentation model, and "**source distillation**" refers to replacing real images with synthetic images for distillation. For all methods, the DeepLab network is deployed due to its efficiency. All other experimental settings are identical with previous experiments.

The results of different methods are summarized in Fig 4 where the "NonAdapt" baseline is also included for comparison. From the figure, we observe that all methods achieve performance gain when compared with the "NonAdapt" baseline, which implies that it is the beneficial to prevent the model from overfitting the synthetic images. Directly freezing a few layers is similar to the "source distillation' method. Our target guided distillation method
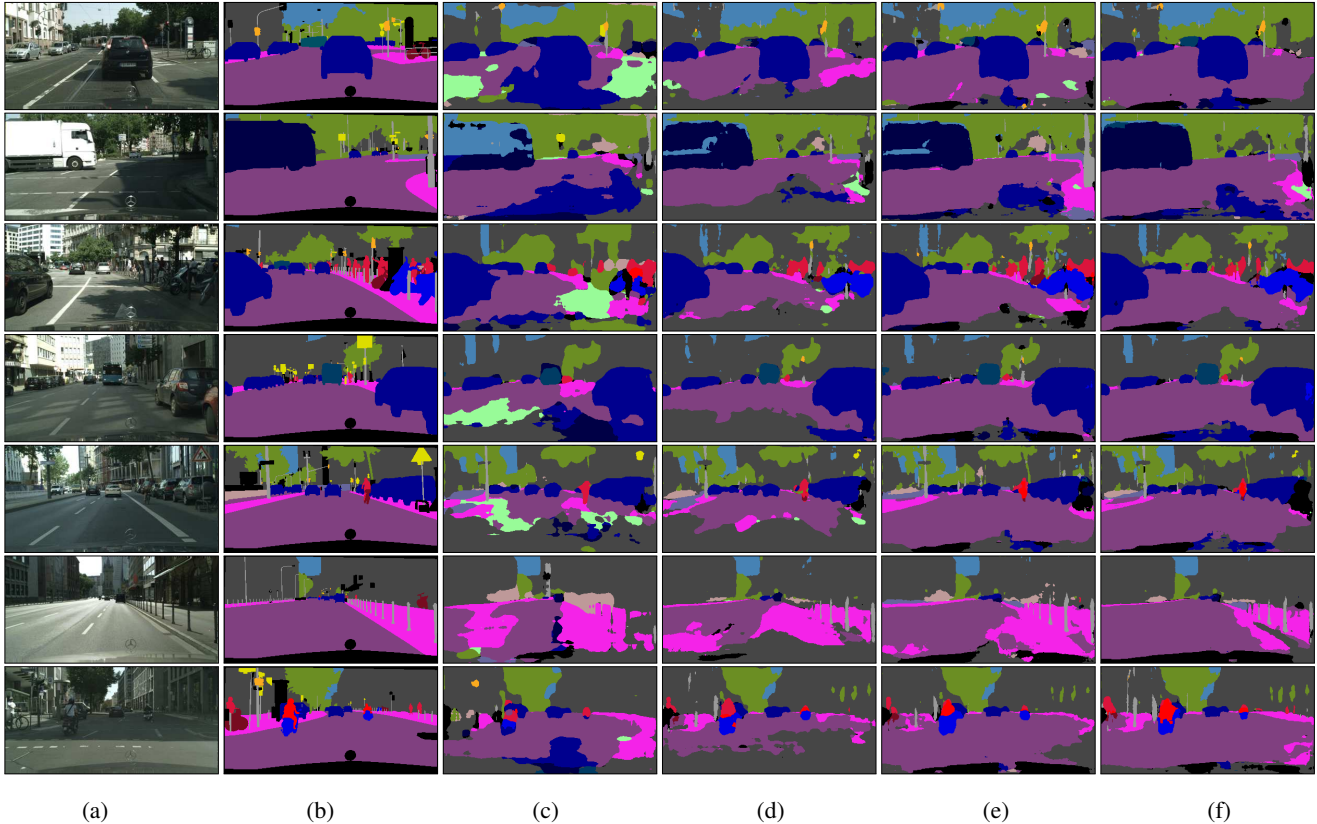
Figure 5. Qualitative semantic segmentation results on the Cityscapes: (a) original image, (b) ground truth annotation, (c) NonAdapt baseline, (d) our model using target guided distillation only, (e) our model using spatial aware adaptation only, (f) our final ROAD-Net model.
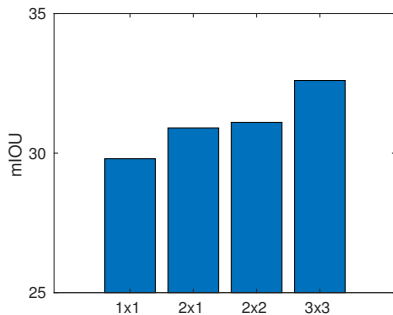

Figure 6. Effects of different partitions in spatial-aware adaptation. $H \times W$ denotes that the image is divided H times along the vertical axis, and W times along the horizontal axis.

achieves the best improvement, which demonstrates its effectiveness by using real images in distillation to prevent the models from over-fitting synthetic style images.

4.4. Analysis on Real Distribution Orientation

In our experiments, we divide the image into 3×3 spatial regions for our spatial-aware adaptation module. To study the effect of different partitions of regions on our method, we conduct additional experiments by using 1×1, 2×1,

2×2 partitions, respectively. The DeepLab network is used in this experiment, while other experimental settings remain unchanged with previous experiments.

The results by using different region partitions are shown in Fig. 6. Among those, the case of 1×1 is the ordinary domain adaptation method without considering the spatial information, which is also equivalent to the *global alignment* method in [17]. We observe that all other partition methods outperform this baseline, which demonstrate the benefit of exploiting spatial information when aligning the domain distribution for urban scenes. The 2×1 (*i.e.*, a top region and a bottom region) gains improvement when compared to the baseline, largely because it uses the semantic prior(*e.g.*, the *sky* is in the top region whereas the *road* is in the bottom region). 2×2 gets little further improvement over 2×1, mainly because that the urban scene is generally vertically symmetric, thus a further partitioning the regions along the vertical middle does not help too much. The 3×3 partition gives the best results, a possible explanation is that it contains a central region, which further copes with the size variance caused by the perspective transformation (*e.g.*, the objects in the central region are smaller, while the objects outside the central are larger).

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mean IoU
GTAV NonAdapt [17]	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	21.1
FCNs Wld [17]	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
NonAdapt [49]	18.1	6.8	64.1	7.3	8.7	21.0	14.9	**16.8**	45.9	2.4	64.4	41.6	**17.5**	55.3	8.4	5.0	6.9	4.3	13.8	22.3
Curriculum [49]	74.8	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**	**14.6**	28.9
NonAdapt	29.8	16.0	56.6	9.2	17.3	13.5	13.6	9.8	74.9	6.7	54.3	41.9	2.9	45.0	3.3	13.1	1.3	6.8	0.0	21.9
Ours	**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**
SYNTHIA NonAdapt [17]	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
FCNs Wld [17]	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
NonAdapt [49]	5.6	11.2	59.6	0.8	**0.5**	21.5	8.0	5.3	72.4	-	75.6	35.1	9.0	23.6	-	4.5	-	0.5	18.0	22.0
Curriculum [49]	65.2	26.1	74.9	0.1	**0.5**	10.7	3.7	3.0	76.1	-	70.6	47.1	8.2	43.2	-	**20.7**	-	0.7	13.1	29.0
NonAdapt	4.7	11.6	62.3	**10.7**	0.0	22.8	4.3	15.3	68.0	-	70.8	**49.7**	6.4	60.5	-	11.8	-	2.6	4.3	25.4
Ours	**77.7**	**30.0**	**77.5**	9.6	0.3	**25.8**	**10.3**	15.6	**77.6**	-	**79.8**	44.5	**16.6**	67.8	-	14.5	-	**7.0**	**23.8**	**36.2**

Table 2. Comparison with state-of-the-arts methods for semantic segmentation on Cityscapes using synthetic datasets as the training data. **Top:** adapting from GTAV, **Bottom:** adapting from SYNTHIA. Results of state-of-the-art methods are from their papers. We use VGG-16 as the backbone network for fair comparison. The best results are denoted in bold.

4.5. Comparison to State-of-the-arts

In this section, we compare our method with two recent methods on semantic segmentation of urban scenes. The first one is *FCNs in the Wild* [17], which also adopted an adversarial training strategy, and further deployed a category specific adaptation method to align pixels of two domains that are likely from the same category. The second one is the *curriculum learning adaptation* [49], in which a curriculum learning approach was used to progressively adapt to the target domain. In both works, the dilation model with VGG-16 backbone was used as the base network.

For a fair comparison, we compare the results of our ROAD-Net based on VGG-16 model with those two state-of-the-arts, which are summarized in the top part of Table 2. The "NonAdapt" baselines for each work are also included for comparison. We observe that, the results from our "NonAdapt" baseline is similar to those reported in [17, 49], despite that we use the DeepLab model while Dilation network is used their experiment. Moreover, by using the proposed two module for real style orientation and real distribution orientation, our ROAD-Net achieves a mean IoU of 35.9%, which outperforms the two state-of-the-art methods. Specifically, we achieves +7.0% and +8.8% improvement compared to [49] and [17] respectively.

4.6. Additional Results on SYNTHIA

To further validate the effectiveness of the proposed method, we additionally perform an experiment using SYNTHIA [33] as source domain and Cityscapes as target domain. SYNTHIA is a dataset with synthetic images of urban scenes, with pixel-wise annotations. The rendering are across a variety of environments and weather conditions. In our experiment, we use the SYNTHIA-RAND-CITYSCAPES subset, which contains 9,400 images compatible with the cityscapes classes. All experiment settings remain unchanged with the previous experiments. For fair comparison, VGG-16 is used as backbone model. The results of all methods are summarized in the bottom part of Table 2. The "NonAdapt" baseline for each method is also included for comparison. We observe that our proposed approach outperforms the other methods by a large margin, which again demonstrates the effectiveness of our ROAD Net in cross-domain semantic segmentation of urban scenes.

5. Conclusion

In this paper, we have presented a new model *Reality Oriented ADaptation Network*(ROAD-Net) for semantic segmentation of urban scenes by learning from synthetic data. Two modules are proposed in our paper, *target guided distillation* and *spatial-aware adaptation*, where the former one aims to adapt the style from the real images by imitating the pretrained network, and the latter one is used to reduce the domain distribution mismatch with the help of layout information. Those two modules can be integrated with different semantic segmentation networks to improve their generalizability when applying to a new domain. We evaluate the proposed method on Cityscapes, using GTAV and STYNHIA as the source domain. The experiments on benchmark datasets have clearly demonstrated the effectiveness of our proposed ROAD-Net.

References

[1] 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. *arXiv:1606.00915*, 2016. 2, 3, 5

[2] Y. Chen, D. Dai, J. Pont-Tuset, and L. Van Gool. Scale-aware alignment of hierarchical image segmentation. In *CVPR*, 2016. 4

[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive Faster R-CNN for object detection in the wild. In *CVPR*, 2018. 3

[4] 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 *ICCV*, 2017. 3

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 3

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 2

[8] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2, 4, 5

[9] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, 2016. 2

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 5

[12] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 2

[13] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 3

[14] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers. Associative domain adaptation. In *ICCV*, 2017. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 3

[17] J. Hoffman, D. Wang, F. Yu, and T. Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:1612.02649*, 2016. 1, 2, 4, 5, 6, 7, 8

[18] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008. 3

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[20] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 2

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 3

[22] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 2

[23] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool. Domain generalization and adaptation using low rank exemplar SVMs. *T-PAMI*, 2017. 2

[24] Z. Li and D. Hoiem. Learning without forgetting. In *ECCV*, 2016. 3, 4

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 5

[26] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[27] H. Lu, L. Zhang, Z. Cao, W. Wei, K. Xian, C. Shen, and A. van den Hengel. When unsupervised domain adaptation meets tensor representations. In *ICCV*, 2017. 2

[28] F. Maria Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. Rota Bulo. Autodial: Automatic domain alignment layers. In *ICCV*, 2017. 2

[29] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017. 2

[30] P. Panareda Busto and J. Gall. Open set domain adaptation. In *ICCV*, 2017. 2

[31] X. Peng and K. Saenko. Synthetic to real adaptation with deep generative correlation alignment networks. *arXiv:1701.05524*, 2017. 3

[32] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 1, 5

[33] 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 *CVPR*, 2016. 3, 8

[34] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv:1712.02560*, 2017. 3

[35] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *arXiv:1711.06969*, 2017. 3

[36] O. Sener, H. O. Song, A. Saxena, and S. Savarese. Learning transferrable representations for unsupervised domain adaptation. In *NIPS*, 2016. 2

[37] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 2

[38] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *arXiv:1612.07828*, 2016. 3

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 3

[40] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. *arXiv:1511.05547*, 2015. 2

[41] B. Sun and K. Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, 2014. 3

[42] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 2

[43] 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. *arXiv:1802.10349*, 2018. 3

[44] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo. Virtual and real world adaptation for pedestrian detection. *T-PAMI*, 36(4):797–809, 2014. 3

[45] L. Wang, Z. Wang, Y. Qiao, and L. Van Gool. Transferring deep object and scene representations for event recognition in still images. *IJCV*, 2017. 3

[46] J. Xu, S. Ramos, D. Vázquez, and A. M. López. Hierarchical adaptive structural svm for domain adaptation. *IJCV*, 2016. 3

[47] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015. 2, 3, 5

[48] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. In *ECCV*, 2010. 2

[49] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. *arXiv:1707.09465*, 2017. 1, 2, 3, 5, 6, 8

[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv:1612.01105*, 2016. 2, 3, 5