# Weakly-supervised Cross-domain Road Scene Segmentation via Multi-level Curriculum Adaptation

Fengmao Lv, Guosheng Lin, Peng Liu, Guowu Yang,
Sinno Jialin Pan, and Lixin Duan

***Abstract*—Semantic segmentation, which aims to acquire pixel-level understanding about images, is among the key components in computer vision. To train a good segmentation model for real-world images, it usually requires a huge amount of time and labor effort to obtain sufficient pixel-level annotations of real-world images beforehand. To get rid of such a nontrivial burden, one can use simulators to automatically generate synthetic images that inherently contain full pixel-level annotations and use them to train a segmentation model for the real-world images. However, training with synthetic images usually cannot lead to good performance due to the domain difference between the synthetic images (i.e., source domain) and the real-world images (i.e., target domain). To deal with this issue, a number of unsupervised domain adaptation (UDA) approaches have been proposed, where no labeled real-world images are available. Different from those methods, in this work, we conduct a pioneer attempt by using easy-to-collect image-level annotations for target images to improve the performance of cross-domain segmentation. Specifically, we leverage those image-level annotations to construct curriculums for the domain adaptation problem. The curriculums describe multi-level properties of the target domain, including label distributions over full images, local regions and single pixels. Since image annotations are "weak" labels compared to pixel annotations for segmentation, we coin this new problem as weakly-supervised cross-domain segmentation. Comprehensive experiments on the `GTA5` → `Cityscapes` and `SYNTHIA` → `Cityscapes` settings demonstrate the effectiveness of our method over the existing state-of-the-art baselines.**

***Index Terms*—Semantic segmentation, domain adaptation, weakly-supervised learning.**

## I. Introduction

Semantic segmentation, which aims to acquire detailed understanding about images, is an essential problem in computer vision. Different from image recognition, it requires to predict the meaning of each pixel, leading to significantly harder challenges. In this paper, we mainly focus on semantic segmentation in road scenes.

Although deep neural networks have gained great advances in semantic segmentation over the past years [1], such performance leaps are partly at the time and economic cost of huge amounts of human annotations for images, especially for road scene segmentation, which requires to annotate the image pixels of diverse categories in a single image. Such heavy burdens usually become the main bottleneck for training a good segmentation model.

In order to avoid such pixel-level annotation burdens in semantic segmentation, very recently researchers have looked into using weak labels, such as image labels [2], bounding boxes [3], points [4] or scribbles [5]. Compared to pixel-level labels, these weak labels are much easier to be collected. However, prior weakly-supervised learning approaches only work on segmenting salient foregrounds from simple scenes. For real-world road scenes featured by complicated environments, diverse categories and occlusion, how to effectively leverage the image-level labels for semantic segmentation remains unclear.

Besides weakly-supervised learning, training deep models with synthetic images, which are obtained from game simulators (e.g., Grand Theft Auto V) [6], [7], is becoming an alternative to ease the labelling efforts in semantic segmentation. Specifically, these synthetic images simulate real road scenes in cities and their pixel-level annotations can be automatically generated. But due to the noticeable difference in visual effect between the synthetic and real-world images (e.g., coloring, lighting, appearance, etc.), there exists a domain gap which may cause a significant performance drop for real-world image segmentation. To reduce the domain gap, the aforementioned works proposed various domain adaptation algorithms to better adapt from synthetic images to real-world ones, including domain adversarial methods [8], [9], [10] and curriculum learning methods [11], [12]. Existing domain adversarial methods mainly focus on aligning the features of pixels and ignore the structural layout of images [8], [9], [10]. Although the recently proposed curriculum learning methods [11], [12] take layout information into consideration, the constructed curriculum may be incorrect since supervision from the target domain is entirely not provided. The incorrect curriculums can cause negative transfer, and hence degrade the performance of target domain.

Considering the limitations of the previous unsupervised domain adaptation approaches, in this work, we conduct a pioneer attempt of introducing image-level labels in cross-domain segmentation. Although it is very costly to acquire pixel-level annotations, obtaining image-level labels is much easier. Therefore, we assume that image-level labels are available for real-world images in the target domain (see Fig. 1). And we coin this new setting as *weakly-supervised cross-domain*

F. Lv is with Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Sichuan 610074, China. E-mail: (fengmaolv@126.com).

G. Lin and S. J. Pan are with School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore. E-mail: (gslin@ntu.edu.sg, sinnopan@ntu.edu.sg).

P. Liu, G. Yang and L, Duan are with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Sichuan 611731, China. E-mail: (liupeng@std.uestc.edu.cn, guowu@uestc.edu.cn, lxduan@gmail.com).

*segmentation* or *weakly-supervised domain adaptation for segmentation*. This assumption is reasonable as it is easier to collect image-level annotations than instance-level annotations from existing datasets or an image search engine. Specifically, for a single image of the Cityscapes dataset, the pixel-level annotation takes more than 1.5 hour, while the image-level annotation can be finished within one minute. To the best of our knowledge, our work is *a very pioneer one* to tackle this setting for semantic segmentation in road scenes.

To tackle this problem, we propose a novel method dubbed Weakly-supervised Multi-level Curriculum Adaptation (WsMCA). To be specific, the curriculums refer to multi-level properties about target domain, including label distributions over full images, local regions and pixels [11]. The label distributions over pixels can be considered as the finest-grained curriculum. These curriculums constitute the ingredients for adapting the segmentation network to the target domain. Similar to [12], WsMCA constructs the curriculums by exploring the segmentation network itself at each iteration and then uses them to update its parameters. However, unlike the current curriculum adaptation methods [11], [12], WsMCA constructs the fine-grained curriculum on the basis of the coarse-grained one, and hence produces more reliable curriculums. Fig. 2 displays the overall architecture of WsMCA. In order to fully leverage the weak supervision from target domain, WsMCA constitutes multiple curriculums for adapting the label distributions of target images at different levels: 1) overview of full images; 2) local regions of foregrounds or backgrounds; 3) possible pixels for each category. These multi-level curriculums can reveal complementary properties of target images, and hence lead to better guidance for domain adaptation. Extensive experiments clearly demonstrate the effectiveness of our method for leveraging weak supervision in cross-domain segmentation. We believe that this work itself can be a strong baseline for cross-domain weakly supervised semantic segmentation.

The main contributions of this work are as follows:

- We conduct a pioneer attempt of introducing image-level labels, which are much easier to be collected than the pixel-level labels, to improve the performance of cross-domain segmentation.
- We propose the Weakly-supervised Multi-level Curriculum Adaptation method to construct diverse reliable curriculums for advanced domain adaptation.
- Extensive experiments are conducted to demonstrate the effectiveness of our method for leveraging weak supervision in cross-domain segmentation.

This paper is organized as follows. Section II reviews the related literatures. Section III introduces the problem statement and the motivation. Section IV presents the proposed method for weakly-supervised cross-domain segmentation. Section V displays the experimental results on several standard benchmarks. Finally, Section VI summarizes this paper.

## II. Related Work

### A. Semantic Segmentation

Semantic segmentation in images is a very important research topic in computer vision. With the recent advances of Fully Convolutional Networks (FCN) [13], various clever techniques, raining from to multi-scale aggregation [14], [15], [16] to context relation [17], [18], were proposed to design advanced segmentation networks. Besides, post processing techniques like conditional random fields [19] were also exploited to improve the performance of semantic segmentation neural networks. However, training segmentation networks requires a huge amount of time and labor effort to acquire sufficient pixel-level annotations of real-world images beforehand.

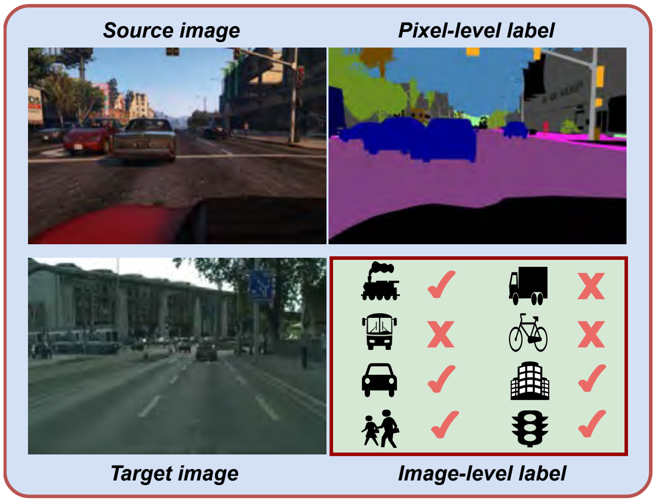


Fig. 1: Illustration of the weakly-supervised cross-domain segmentation problem. Source images (i.e., synthetic images) have full pixel-level annotations, while only image-level class labels are available for target images (i.e., real-world images).

### B. Weakly-supervised Semantic Segmentation

Weakly-supervised segmentation aims to alleviate the workload of pixel-wise labels for the training data through leveraging weak labels, such as image labels that would be far cheaper to be collected. Weakly-supervised segmentation with image-level labels was originally tackled through Multiple Instance Learning (MIL) [20] or Expectation-Maximization (EM) mechanism [21]. The recent works primarily focused on guiding top-down segmentation cues in fully convolutional networks with image-level labels [2], [22], [23], [24]. These cues visually reflect class-specific cues on both localization and objection size. Besides image-level labels, other weak supervision, such as bounding boxes [3], points [4] or scribbles [5], were also exploited to implement weakly-supervised semantic segmentation. However, the current weakly-supervised learning methods were mainly designed for relatively simple tasks segmenting the salient foreground objects in each single image [2], [20], [25]. To the best of our knowledge, no works have effectively leveraged the image-level labels in road scene segmentation featured by complicated environments and diverse categories of both foregrounds and backgrounds.

### C. Domain Adaptation for Semantic Segmentation

Most of the previous works on visual domain adaptation focused on the classification task. Basically, the main idea is

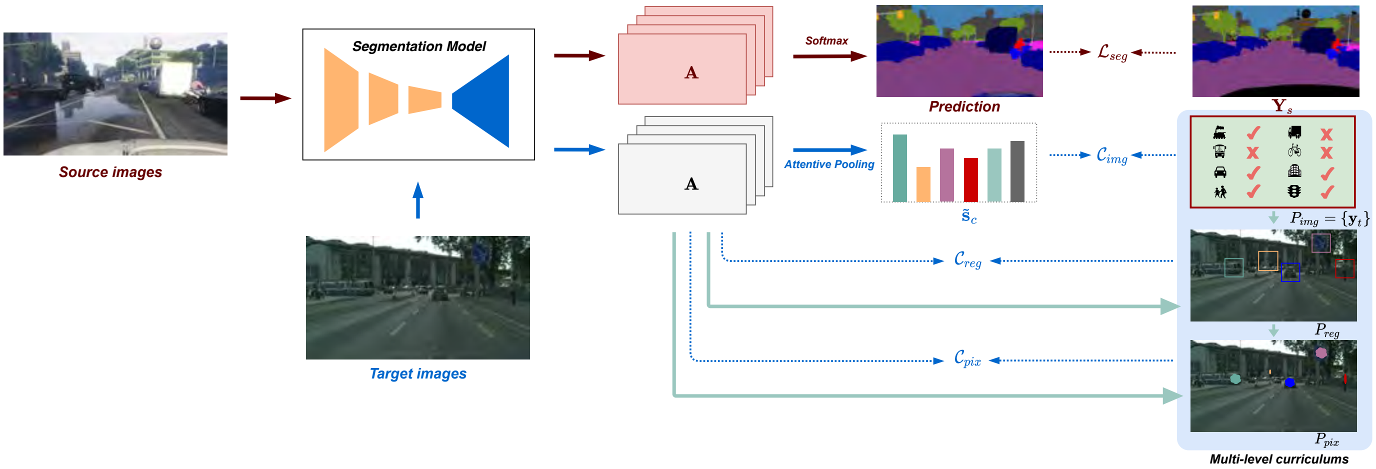


Fig. 2: The overall architecture of our proposed WsMCA (best viewed in color). WsMCA constructs curriculums over full images, local regions and pixels to adapt the semantic segmentation networks into the target domain. The fine-grained curriculum is generated on the basis of the coarse-grained one. The green lines depict the flow paths of constructing curriculums.

to bridge the source and the target domains by distribution alignment or pseudo labels [26], [27], [28], [29], [30]. Similar to image classification, one of the primary approaches to cross-domain semantic segmentation is to align the distributions through enforcing domain adversarial training over the intermediate representations [8], [9], [10], [31], [32], [33], [34], [35] or the structured output space [36], [43] of semantic segmentation neural networks. In [38] and [39], Generative Adversarial Networks (GANs) were adopted to directly translate the source images into target-style images, which were then used to train segmentation networks for target domain. Recently, curriculum learning methods were also designed to adapt segmentation networks through constructing curriculums that reveal the labelling distributions of target images [11], [12], [40]. However, the constructed curriculums may be incorrect since supervision from the target domain is entirely not provided.

Besides strategies over the methodology aspect, internal data are also leveraged to improve the performance of cross-domain semantic segmentation. Lin et al. proposed to reduce domain variations in multi-person part segmentation by using pose labels from the target domain [41]. Lee et al. and Vu et al. proposed to use the dense depth from the source domain as privileged information [42], [43]. Wang et al. proposed to use the bounding box annotations from the target domain to improve the performance [44]. However, since the semantic segmentation task for real road scenes is featured by several thorns (e.g., diverse categories, occlusion of backgrounds, multiple objects with different scales), collecting weak supervision of bounding boxes from target domain is costly.

## III. Problem Statement & Motivation

Formally, in the setting of this work, we are given source images $\mathbf{I}_s \in \mathbb{R}^{H\times W\times 3}$ with pixel labels $\mathbf{Y}_s \in \{0,1\}^{H\times W\times C}$ and target images $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ with image-level labels $\mathbf{y}_t \in \{0,1\}^{C}$. To be specific, $H$ and $W$ respectively denote the height and width of images, while $C$ denotes the category number. Denote by $\mathcal{S}$ and $\mathcal{T}$ the source domain and the target domain, respectively. Our goal is to learn a good segmentation network that can achieve desirable pixel-level prediction performance over the target domain.

Our work is mainly motivated by the limitations of the current curriculum methods for cross-domain segmentation [11], [12]. Their main idea is to use additional models or the segmentation network itself to construct curriculums for domain adaptation. However, as supervision from the target domain is entirely not provided, the constructed curriculums are at high risk of being inconsistent with the ground truth. The incorrect curriculums will guide segmentation networks to learn noisy information, and hence lead to negative transfer. Therefore, this work proposes a multi-level curriculum adaptation approach to construct more reliable curriculums for cross-domain semantic segmentation. To this end, we introduce image-level labels from real road scenes as the coarsest-grained curriculum. The finer-grained curriculum is then generated on the basis of the coarse-grained one. As displayed in Fig. 2, the curriculums include structural layout of full images, local regions of foregrounds (or backgrounds) and possible pixels for each class. The curriculum over each level plays a unique role for revealing the properties of target domain. Specifically, the curriculum over full images maintains the overview of target images, and the curriculum over local regions reveals specific locations for each category. The pixel-level curriculum further provides more precise supervision, which is a necessity for pixel-level prediction. These complementary curriculums can jointly give better guidance for domain adaptation.

## IV. Weakly-supervised Multi-level Curriculum Adaptation

In this section, WsMCA is presented for weakly-supervised domain adaptation for segmentation. To fully leverage the weak supervision from target domain, we propose to construct multi-level curriculums for revealing the target properties and using the properties for adapting segmentation network towards target domain.

**Preliminary: curriculum domain adaptation**. Constructing curriculums for target domain is an alternative towards domain adaptation for semantic segmentation. Curriculum means to obtain easily available knowledge about the target domain first and then use it to regularize the network predictions [11]. The knowledge is about the desired properties that should be met by the outputs of semantic segmentation neural networks. Compared to pseudo labels [12], such the properties can lie different aspects. Curriculum methods achieve domain adaptation through first inferring desired properties about the label distribution of target images, $P(\mathbf{I}_t)$, and then enforcing segmentation network's softmax prediction on target images, $S(\mathbf{I}_t) \in \mathbb{R}^{H \times W \times C}$, to meet these properties [11]:

$$\min \sum_{s \in \mathcal{S}} \mathcal{L}_{seg}(S(\mathbf{I}_s), \mathbf{Y}_s) + \sum_{t \in \mathcal{T}} \mathcal{C}(S(\mathbf{I}_t), P(\mathbf{I}_t)), \quad (1)$$

where $\mathcal{L}_{seg}$ denotes the pixel-wise cross-entropy loss and $\mathcal{C}$ denotes the curriculum loss for meeting properties $P(\mathbf{I}_t)$. The target properties can be constructed by either additional models [11] or the segmentation networks itself [12].

**Overview**. In this work, curriculums are designed over three aspects, including full image, local region and single pixel, which are discussed in the following. Specifically, the curriculum over images refers to properties in the form of image-level labels, i.e., which classes are included in a target image, and the curriculum over local regions refers to properties in the form of label positions, i.e., where each class may exist in a target image. Pseudo labels can be considered as the finest-grained property about the target domain. Hence, the curriculum loss in Eq. (1) inculdes three terms:

$$\mathcal{C} = \lambda_{img}\mathcal{C}_{img} + \lambda_{reg}\mathcal{C}_{reg} + \lambda_{pix}\mathcal{C}_{pix},$$

where $\lambda_{adv}$, $\lambda_{img}$, $\lambda_{reg}$ and $\lambda_{pix}$ are the trade-off parameters that weigh the importance of the corresponding terms. In order to tackle the setting of weak-supervised domain adaptation, our method proposes to constitute multi-level curriculums for adapting the label distributions of target images over different views. Specifically, the curriculums include overview of full images, local regions of foregrounds (or backgrounds) and possible pixels for each category. For updating the parameters of $S$, we will resort to the segmentation network itself for generating the curriculums at each iteration, and then use the current curriculums to obtain the gradients for update.

Moreover, we also incorporate the common-used domain-adversarial loss [36] in our objective. To this end, a discriminative network $D$ is introduced to perform adversarial training over the output space of the segmentation network $S$ to reduce the distribution shift. Denote by $\mathcal{L}_{adv}$ the domain-adversarial term and $\lambda_{adv}$ the corresponding trade-off parameter. The overall objective is formulated as follows:

$$\max_{D} \min_{S} \ \mathcal{L}_{seg} + \mathcal{C} + \lambda_{adv}\mathcal{L}_{adv}.$$

We refer the readers to [36] for more details about the $\mathcal{L}_{adv}$ loss. The curriculum loss terms are formulated as follows.

**Full images**. The curriculums designed over full images are for revealing the global label distribution of target images. To this end, we directly employ the image-level label from target domain, $\mathbf{y}_t$, to form curriculum over full images:

$$P_{img}(\mathbf{I}_t) = \{\mathbf{y}_t\}.$$

To be specific, image-level labels from target domain tell us which objects or backgrounds should exist in a target image.

To enforce the output of semantic segmentation neural networks to meet the property of $P_{img}(\mathbf{I}_t)$, a natural idea is to perform Multiple Instance Learning over target images. However, for each category, MIL only encourages one single pixel to have high activation value, which is relatively weak for road scenes featured by complicated environments. For expanding the object areas of each category, we design an attentive learning module. This module consists of a dropout layer, an attentive pooling layer and a multi-label classification loss layer. Firstly, we incorporate a dropout layer before producing class activation maps, which helps to explore more category relevant regions instead of only focusing on the most relevant position. Then, following [45], we conduct attentive pooling over the activation maps, denoted by $\mathbf{A} \in \mathbb{R}^{H \times W \times C}$, which are the network layer outputs right before the segmentation softmax layer. This attentive pooling can be written as:

$$s^{(c)} = \log \left[ \frac{1}{HW} \sum_{h,w} \exp(r \cdot \mathbf{A}^{(h,w,c)}) \right]. \quad (2)$$

The pooling output value is denoted by $s^{(c)}$. In the above equation, the exponential operation suppresses pixels with small activation values, while retains pixels with large activation values. In principle, the attentive pooling in Eq. 2 acts as a soft version of the max pooling layer in MIL, and it results in the effects that the image regions containing particular objects will have high activation values on the activation map – the target object regions will be highlighted on the activation map. The hyper-parameter $r$ in Eq. 2 acts as a smooth parameter that controls how smoothly the activation value contributes to the pooling output value $s^{(c)}$. Specifically, the attentive pooling with a large value of $r$ tends to select the regions with large activation values to generate the output, while the one with a small value of $r$ tends to equally consider every spatial position in the activation map to generate the output. In our work, we simply fix the smooth parameter $r$ to 1, which is sufficient to achieve good object localization performance. With the pooling output value $s^{(c)}$, we then perform image-level multi-label classification:

$$\mathcal{C}_{img}(\mathbf{y}_t) = \sum_{c} \left[ (y_{t,c} - 1) \log \frac{e^{-s^{(c)}}}{1 + e^{-s^{(c)}}} - y_{t,c} \log \frac{1}{1 + e^{-s^{(c)}}} \right].$$

The module can drive the segmentation network to highlight target object regions – the image regions containing the objects corresponding to $\mathbf{y}_t$. Hereinafter, $\mathcal{C}_{img}(\mathbf{y}_t)$ is denoted as $\mathcal{C}_{img}$.

**Local regions**. The curriculums designed over local regions are for revealing specific locations for each category. As the same as curriculums designed over full images, the curriculums over local regions are also constructed under the guidance of the weak supervision $\mathbf{y}_t$ from target domain. To this end, we generate category-specific patches to form the

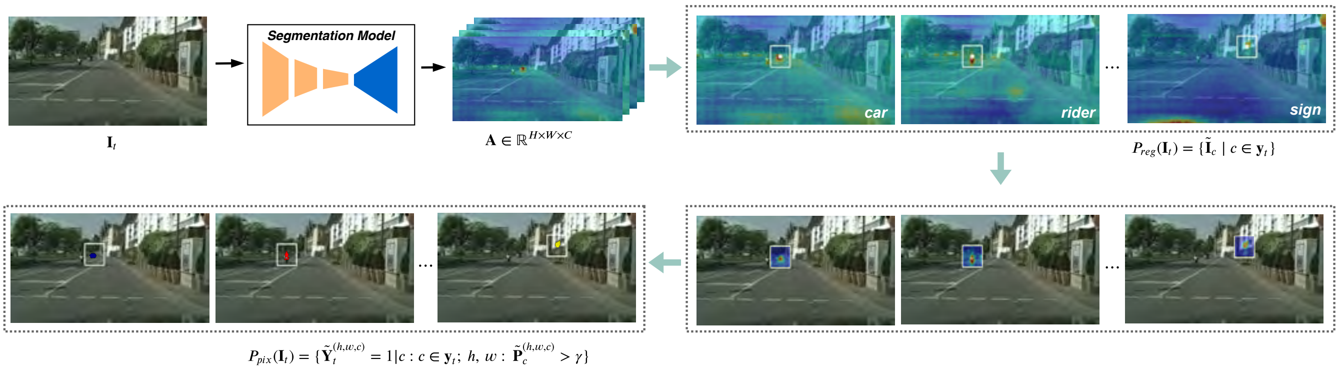

Segmentation Model
$\mathbf{I}_t$
$\mathbf{A} \in \mathbb{R}^{H\times W\times C}$
car
rider
sign
...
$P_{reg}(\mathbf{I}_t) = \{\tilde{\mathbf{I}}_c \mid c \in \mathbf{y}_t\}$
$P_{pix}(\mathbf{I}_t) = \{\tilde{\mathbf{Y}}_t^{(h,w,c)} = 1 | c : c \in \mathbf{y}_t;\ h,\ w : \tilde{\mathbf{P}}_c^{(h,w,c)} > \gamma\}$

Fig. 3: The process to generate curriculums over local regions and image pixels. The green lines depict the flow paths of constructing curriculums. For each category that exists in a target image, we first locate the pixel position that has the largest value in the corresponding class activation map and then generate a fixed-size patch centered at that position as the curriculums over local regions. The pixel-level curriculums are generated based on the curriculums over local regions. Within the region patch $\tilde{\mathbf{I}}_c$, we assign pseudo labels to pixels that have high confidence in being class $c$. Compared to [12], our approach only assigns pseudo labels to pixels that sit inside the target object regions.

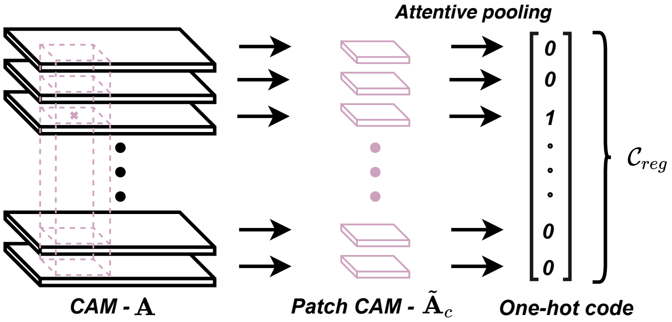


Fig. 4: Attentive learning module over local regions. The cross denotes the pixel that maximally activates the $c$-th category: $\mathrm{argmax}_{\mathrm{h,w}}\ \mathbf{A}^{(\mathrm{h,w,c})}$. We generate a fixed-size patch around that pixel. Over the generated patch activation map, softmax classification is performed.

curriculums. Similar to [12], these curriculums are constructed through resorting to the segmentation network itself. To be specific, for a category that appears in the target image, we identify the pixel position which maximally activates that category, and then generate a fixed-size patch centered at that position for this particular category. This process is illustrated in Fig. 3. Remarkably, we also include the dropout layer before generating the patch activation map to encourage the algorithm to generate curriculums over more diverse regions rather than always selecting a particular position. Denoting these local regions by $\tilde{\mathbf{I}}_c$, where $c \in \mathbf{y}_t$, we can represent the curriculums over local regions as follows:

$$P_{reg}(\mathbf{I}_t) = \{\tilde{\mathbf{I}}_c \mid c \in \mathbf{y}_t\}.$$

These generated local regions are assumed to contain the object, or at least the discriminative part, of the corresponding category.

To enforce the output of semantic segmentation neural networks to meet the property of $P_{reg}(\mathbf{I}_t)$, we implement similar attentive learning operation over local regions as that for full images. To be specific, the selected patches go through the segmentation network to obtain the Patch Class Activation Map (PCAM) $\tilde{\mathbf{A}}_c \in \mathbb{R}^{H_0\times W_0\times C}$. Similar to attentive learning module over full images, we apply attentive pooling operation in Eq. (2) on the patch activation map $\tilde{\mathbf{A}}_c$ to generate the pooling outputs $\tilde{\boldsymbol{s}}_c \in \left[\tilde{s}_c^{(1)}, \tilde{s}_c^{(2)}, ..., \tilde{s}_c^{(C)}\right]^T$. The softmax output of the patch activation map $\tilde{\mathbf{A}}_c$ is denoted by $\tilde{\mathbf{P}}_c \in \mathbb{R}^{H_0\times W_0\times C}$. Finally, we apply the following local patch classification loss function:

$$\mathcal{C}_{reg} = -\sum_{c\in\mathbf{y}_t} \log \frac{e^{\tilde{s}_c^{(c)}}}{\sum_{i=1}^{C} e^{\tilde{s}_c^{(i)}}}. \tag{3}$$

This is about the softmax classification with the patch activation map as the input. It is similar to the image-level classification loss for full images. The process is illustrated in Fig. 4. For categories that exist in a target image, $\mathcal{C}_{reg}$ can drive the segmentation network to highlight the corresponding object areas or the discriminative parts in selected local regions, and hence make the segmentation network to observe the property of $P_{reg}(\mathbf{I}_t)$.

**Single pixels**. The curriculums designed over pixels are for providing more precise supervision, which is essential for the pixel-level prediction task. The pixel-level curriculum is also constructed by resorting to the segmentation network itself. However, unlike Zou et al. [12], for a category $c$ that exists in target images, we only select pixels in the corresponding local regions $\tilde{\mathbf{I}}_c$. To be specific, within each local region $\tilde{\mathbf{I}}_c$, we select the pixel positions whose softmax score values for the particular class surpass the threshold $\gamma$. We argue that these selected pixel positions are very likely to sit inside the target object regions. Formally, the curriculum over pixels is represented as follows:

$$P_{pix}(\mathbf{I}_t) = \{\tilde{\mathbf{Y}}_t^{(h,w,c)} = 1 | c : c \in \mathbf{y}_t;\ h,\ w : \tilde{\mathbf{P}}_c^{(h,w,c)} > \gamma\}.$$

The process is illustrated in Fig. 3. As the local regions are very confident areas for the particular category, our design, compared with [12], reduces incorrect pixel-level curriculums and produces more reliable pixel-level curriculums.

To adapt segmentation network to reach property of $P_{pix}(\mathbf{I}_t)$, we can directly select these high-confident positions to perform softmax classification. To be specific, we apply a softmax classification loss for these selected pixel positions with their activation map values as input:

$$\mathcal{C}_{pix} = -\sum_{c \in \mathbf{y}_t} \sum_{h,\, w :\, \tilde{\mathbf{P}}_c^{(h,w,c)} > \gamma} \tilde{\mathbf{Y}}_t^{(h,w,c)} \cdot \log \tilde{\mathbf{P}}_c^{(h,w,c)}. \qquad (4)$$

The symbol $\gamma$ denotes a predefined threshold parameter. This pixel-level curriculum provides more precise supervision for adapting segmentation network into the target domain.

## V. Experiments

### A. Datasets

In our experiments, we conduct thorough evaluations of our proposal on adapting synthetic images to real-world street views, including both GTA5 → Cityscapes and SYNTHIA → Cityscapes.

GTA5 → Cityscapes: The GTA5 dataset includes 24,966 synthetic images with resolution of 1914 × 1052 [50]. Specifically, these synthetic images, simulating the virtual urban views of Los Angeles, are rendered from the game engine of Grand Theft Auto V (GTA5). With computer graphics techniques, the pixel-wise annotations of the GTA5 images can be automatically produced. The Cityscapes dataset is mainly tailored for automatic driving in urban roadways. The images in Cityscapes are real photos featured by scene variability and complexity. In particular, Cityscapes consists of 2,975 training images and 500 validation images. These images have a resolution of 2048 × 1024. In the experiments, 19 common categories among GTA5 and Cityscapes are of interest. We train our model using all the pixel-level annotated GTA5 images and the image-level annotated images from the training set of Cityscapes. Following the existing state-of-the-art works [36], [12], [33], we evaluate our method over the validation set with 500 images.

SYNTHIA → Cityscapes: In this setting, we adopt the SYNTHIA-RAND-CITYSCAPES set as the source domain [7]. In particular, "SYNTHIA" includes 9,400 photo-realistic images rendered from virtual scenes, with the size of 960 × 720. Similar to GTA5 images, their pixel-wise labels are generated automatically. Following [36] and [32], we evaluate our method over 13 common categories. Also, the full "SYNTHIA" dataset is used for training and the data spilt for Cityscapes is identical to the above setting.

### B. Implementation Details

In the experiments, we implement the segmentation network $S$ by FCN8s with VGG-16 [13] and deeplab-v2 with ResNet-101 [14]. The network parameters are pre-trained on ImageNet. Both the topology of the discriminative network $D$ and the value of $\lambda_{adv}$ follow the identical settings in [36].

In our experiments, the GTA5 images and the Cityscapes images are resized to 1280 × 640 and 1024 × 512, respectively, while the resolution of the "SYNTHIA" images are kept unchanged. To validate the robustness of our method, we adopt the identical hyper-parameters in both GTA5 → Cityscapes and SYNTHIA → Cityscapes. Specifically, the parameters of the backbone network $S$ is optimized by stochastic gradient descent (SGD) with momentum of 0.9 and weight decay of 0.0005, while those of the discriminative network $D$ are optimized by Adam with momentum of 0.9 and 0.99. The initial learning rates for $S$ and $D$ are respectively set to 0.00025 and 0.0001. The maximum iteration number is $120k$ and the batch size is set to 2. The curriculums over local regions and fine-grained pixels are progressively incorporated after $8k$ iterations. The trade-off hyper-parameters $\lambda_{img}$ and $\lambda_{reg}(\lambda_{pix})$ are fixed to 0.2 and 0.0005, respectively. Additionally, we fix the size of local patch to 200 × 200. The self-training threshold $\gamma$ in Eq. 4 is set to 0.7. Finally, the mIoU value is adopted as the metric of evaluation.

### C. Baselines

We conduct extensive comparisons with state-of-the-art unsupervised and weakly-supervised domain adaptation methods, which are listed as follows.

**Unsupervised domain adaptation baselines:**

- Source-Only: Semantic segmentation model trained with only the source data. Moreover, we also report the target-only results obtained by the model trained with target data in the supervised setting to serve as the upper bounds of domain adaptation performance.
- DCAN: Dual Channel-wise Alignment Network from [47]. It focuses on matching the channel-wise feature statistics.
- SIBAN: Significance-aware Information Bottlenecked Adversarial Network from [34]. SIBAN purifies features with a significance-aware information bottleneck to stabilize the adversarial training course.
- ROAD: Reality Oriented ADaptation in [46]. It proposes to conduct feature alignment over different regions.
- SSF: Separated Semantic Feature from [35]. SSF aligns the distribution discrepancy via independent class-wise adversarial learning.
- CLAN: Class-Level Adversarial Network in [31]. CLAN adaptively weights the adversarial loss for each feature according to the category-level alignment degree.
- AdaptSeg: Adaptive Segmentation proposed in [36]. It implements domain adaptation through aligning the output space of segmentation networks.
- AdvEnt: Adversarial Entropy minimization approach from [43]. AdvEnt adapts the principle of entropy minimization to cross-domain semantic segmentation.
- DPR: Discriminative Patch Representation from [49]. It proposes to bridge domain shift via patch-level alignment.
- CYCADA: CYCle-consistent Adversarial Domain Adaptation from [39]. CYCADA leverages the recent advances of GANs to directly translate source images to target-style images, which are then used to train segmentation networks for target domain.

TABLE I: Comparisons on `GTA5` → `Cityscapes` in terms of per-class IoUs and mIoU.

| Base Model | Method | road | sdwk | bldng | wall | fence | pole | light | sign | vgttn | trrn | sky | person | rider | car | truck | bus | train | mcycl | bcycl | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGG-16 | Source-Only | 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 |
| | Target-Only | 96.5 | 74.6 | 86.1 | 37.1 | 33.2 | 30.2 | 39.7 | 51.6 | 87.3 | 52.6 | 90.4 | 60.1 | 31.7 | 88.4 | 54.9 | 52.3 | 34.7 | 33.6 | 59.1 | 57.6 |
| | SIBAN [34] | 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 |
| | AdaptSeg [36] | 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 |
| | CyCADA [39] | 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 |
| | ROAD [46] | 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 |
| | DCAN [47] | 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 |
| | CBST [12] | **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 |
| | AdvEnt [43] | 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 |
| | ODC [44] | 85.3 | 43.6 | 78.5 | 28.3 | 25.2 | 10.5 | 10.5 | 6.7 | 81.4 | 33.6 | 74.3 | 36.7 | 3.0 | 73.0 | 20.2 | 13.4 | 0.0 | 4.7 | 0.0 | 33.1 |
| | **WsMCA (ours)** | 85.2 | 26.3 | **78.9** | 19.1 | 22.4 | 19.3 | 18.1 | 18.5 | 80.8 | **34.4** | **79.7** | 47.9 | **20.4** | 78.8 | **29.9** | 22.6 | **21.5** | **18.1** | 2.1 | **38.1** |
| ResNet-101 | Source-Only | 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 |
| | Target-Only | 97.9 | 81.3 | 90.3 | 48.8 | 47.4 | 49.6 | 57.9 | 67.3 | 91.9 | 69.4 | 94.2 | 79.8 | 59.8 | 93.7 | 56.5 | 67.5 | 57.5 | 57.7 | 68.8 | 70.4 |
| | SIBAN [34] | 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 |
| | AdaptSeg [36] | 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 |
| | ROAD [46] | 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 |
| | CLAN [31] | 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 |
| | AdvEnt [43] | 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 |
| | DISE [48] | 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 |
| | DPR [49] | **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 |
| | SSF [35] | 90.3 | 38.9 | 81.7 | 24.8 | 22.9 | 30.5 | 37.0 | 21.2 | **84.8** | 38.8 | 76.9 | 58.8 | 30.7 | 85.7 | 30.6 | 38.1 | 5.9 | 28.3 | 36.9 | 45.4 |
| | **WsMCA (ours)** | 90.3 | 43.4 | 82.4 | 24.1 | **27.1** | 32.9 | **37.8** | **41.0** | 83.2 | 34.1 | 80.8 | 62.0 | **33.7** | 83.3 | **47.7** | **51.3** | **15.0** | **33.8** | **50.3** | **50.2** |

- DISE: Domain Invariant Structure Extraction from [48]. It disentangle images into domain-invariant structure and domain-specific texture representations.
- CDA: Curriculum Domain Adaptation proposed in [11]. It uses additional models to construct curriculums for adapting segmentation networks into target domain.
- CBST: Class-Balanced Self-Training from [12]. CBST constructs curriculums for adaptation through resorting to the segmentation network itself.

**Weakly-supervised domain adaptation baselines:**

- ODC: Object-level Domain Classifier in [44]. It assumes that the weak supervision of bounding boxes from target domain are available for domain adaptation. To leverage the weak supervision of bounding boxes, ODC is designed to learn domain-invariant object features.

Note that collecting weak supervision of bounding boxes in [44] is still costly since the semantic segmentation task for real road scenes is featured by several thorns (e.g., diverse categories, occlusion of backgrounds, multiple objects with different scales).

### *D. Comparisons*

Tables I and II display the comparisons of our method and existing cross-domain semantic segmentation baselines. Compared with the existing UDA baselines for semantic segmentation, it is clear that our proposed multi-level curriculum adaptation methods can fully utilize weak supervision of image-level labels from target domain to improve the performance of cross-domain segmentation significantly. Specifically, the feature alignment approaches, including DCAN, SIBAN, ROAD and SSF, mainly focus on aligning the features of pixels, but ignore the overall structural layout of images. Hence, they achieve relatively poor cross-domain performance. In contrast, the recently proposed structured output adaptation approaches, such as AdaptSeg, AdvEnt and DPR, mainly focuses on adapting the spatial layout of images, but ignores precise adaptation over pixels. The performance of CYCADA and DISE heavily rely on the quality of image translation, which is at discount for the road scene images featured by complicated environments. Both CDA and CBST will suffer from incorrect curriculums since supervision from the target domain is entirely not provided. Noticeably, for ODC with weak supervision of bounding boxes from the target domain that is more costly than image-level labels, its cross-domain performance is relatively poor. This is because that learning domain-invariant object features cannot provide sufficient support for pixel-level prediction, and hence make the weak-supervision of bounding boxes not fully utilized. In contrast, our method constructs multi-level curriculums for revealing diverse properties of target images, which can be used to adapt the network towards better prediction for target domain.

Moreover, in order to verify the effectiveness of our approach in fully utilizing the image-level labels, we replace the proposed multi-level curriculums with other designs, including Multi-Instance Learning (MIL) from [20] and Class Activation Mapping (CAM) from [23]. MIL is the natural approach for leveraging image-level labels. CAM is a common alternative for leveraging image-level labels with global average pooling. From Table III, it is clear that both MIL and CAM only bring very limited performance improvement for cross-domain semantic segmentation. To be specific, for each category,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3040343, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II: Comparisons on `SYNTHIA` → `Cityscapes` in terms of per-class IoUs and mIoU.

| Base Model | Method | road | sdwk | bldng | light | sign | vgttn | sky | person | rider | car | bus | mcycl | bcycl | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGG-16 | Source-Only | 8.4 | 12.0 | 73.4 | 4.5 | 5.2 | 72.8 | 73.5 | 41.3 | 5.2 | 68.8 | 21.9 | 4.9 | 5.6 | 30.6 |
| | Target-Only | 96.5 | 74.6 | 86.1 | 39.7 | 51.6 | 87.3 | 90.4 | 60.1 | 31.7 | 88.4 | 52.3 | 33.6 | 59.1 | 65.5 |
| | CDA [11] | 65.2 | 26.1 | 74.9 | 3.7 | 3.0 | 76.1 | 70.6 | 47.1 | 8.2 | 43.2 | 20.7 | 0.7 | 13.1 | 34.8 |
| | SIBAN [34] | 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 |
| | AdaptSeg [36] | 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 |
| | ROAD [46] | 77.7 | 30.0 | 77.5 | **10.3** | 15.6 | 77.6 | 79.8 | 44.5 | 16.6 | 67.8 | 14.5 | 7.0 | 23.8 | 41.8 |
| | AdvEnt [43] | 67.9 | 29.4 | 71.9 | 0.6 | 2.6 | 74.9 | 74.9 | 35.4 | 9.6 | 67.8 | 21.4 | 4.1 | 15.5 | 36.6 |
| | ODC [44] | **87.4** | **43.4** | 78.0 | 0.0 | 2.9 | **80.1** | **80.5** | 38.1 | 8.1 | 0.0 | **26.2** | 1.4 | 19.7 | 35.8 |
| | **WsMCA (ours)** | 83.5 | 25.4 | 80.7 | 7.9 | **16.2** | 79.2 | 77.5 | **48.0** | **24.0** | **78.8** | 24.1 | **24.7** | **35.8** | **46.6** |
| ResNet-101 | Source-Only | 55.8 | 21.8 | 78.7 | 7.3 | 12.9 | 75.7 | 80.1 | 53.9 | 18.4 | 37.2 | 21.1 | 11.4 | 23.7 | 38.3 |
| | Target-Only | 97.9 | 81.3 | 90.3 | 57.9 | 67.3 | 91.9 | 94.2 | 79.8 | 59.8 | 93.7 | 67.5 | 57.7 | 68.8 | 77.5 |
| | SIBAN [34] | 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 |
| | CLAN [31] | 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 |
| | AdaptSeg [36] | 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 |
| | AdvEnt [43] | 85.6 | 42.2 | 79.7 | 5.4 | 8.1 | 80.4 | 84.1 | 57.9 | 23.8 | 73.3 | 36.4 | 14.2 | 33.0 | 48.0 |
| | DISE [48] | **91.7** | **53.5** | 77.1 | 6.2 | 7.6 | 78.4 | 81.2 | 55.8 | 19.2 | **82.3** | 30.3 | 17.1 | 34.3 | 48.8 |
| | DPR [49] | 82.4 | 38.0 | 78.6 | 3.9 | 11.1 | 75.5 | 84.6 | 53.5 | 21.6 | 71.4 | 32.6 | 19.3 | 31.7 | 46.5 |
| | SSF [35] | 84.6 | 41.7 | 80.8 | 11.5 | 14.7 | 80.8 | **85.3** | 57.5 | 21.6 | 82.0 | 36.0 | 19.3 | **34.5** | 50.0 |
| | **WsMCA (ours)** | 85.8 | 42.6 | **83.7** | **25.3** | **17.1** | **83.1** | 85.0 | 57.1 | **33.0** | 77.6 | **46.1** | **31.3** | 32.4 | **53.9** |

TABLE III: Comparisons of different designs to utilize the image-level labels for pixel-level transfer.

| Method | Base Model | `GTA5` ↓ `Cityscapes` | `SYNTHIA` ↓ `Cityscapes` |
|---|---|---|---|
| CAM [23] | VGG-16 | 32.5 | 39.8 |
| MIL [20] | VGG-16 | 36.8 | 39.2 |
| WsMCA (ours) | VGG-16 | **38.1** | **46.6** |
| CAM [23] | ResNet-101 | 43.4 | 48.5 |
| MIL [20] | ResNet-101 | 43.6 | 47.4 |
| WsMCA (ours) | ResNet-101 | **50.2** | **53.9** |

MIL only encourages one single pixel to have high activation value, which is relatively weak for road scenes featured by complicated environments. On the contrary, for categories that appear in a target image, CAM encourages all pixels to have high activations and overestimates the corresponding region size. Due to the large scale discrepancy among categories in road scenes, CAM may even cause negative transfer.

### *E. Analysis*

**Ablation study**. To further evaluate the contribution of each design in our method, we display the ablation study results in Table IV. The first row represents the segmentation network trained with only the source domain. The second row shows the results of the domain adversarial training baseline. From the third row, it is clear that the curriculum over full images makes a good contribution for adapting segmentation network into the target domain, especially for the infrequent categories (e.g., train, motorcycle, truck and rider). The fourth and the fifth rows progressively incorporate curriculums over local regions and fine-grained pixels. It is clear that both of them provide more precise supervision for domain adaptation, especially for the small-scale categories (e.g., light, sign, bicycle and pole). Notably, by comparing the second and the last row, we can see that the image-level supervision from target domain makes a great contribution for adapting segmentation network towards better prediction over target data.

**Sensitivity analysis**. In Table V, we display sensitivity analysis results. For each design, it is clear that the result is not sensitive to the values of the corresponding hyper-parameters. This demonstrates that our method can tolerate a wide range of hyper-parameters.

**Qualitative results**. For a qualitative view, we illustrate the segmentation results in Fig. 5. In general, leveraging the image-level labels significantly improves the pixel-level prediction of target images, especially for the small-scale foreground objects that are invisible in complicated road scenes. Furthermore, we visualize the class activation maps of CNNs in Fig. 6. As we can see, adapted by curriculums designed with image-level supervision, the segmentation network can produce activation maps which are able to highlight those object regions corresponding to the image-level category labels. As displayed, when curriculums with image-level supervision are removed, the activation maps cannot clearly reveal the object areas for particular categories.

## VI. Conclusion

In this work, we conduct a pioneer attempt to leverage the easy-to-collect image-level annotations for target images to improve the performance of cross-domain segmentation. We coin this new setting as weakly-supervised cross-domain segmentation. To fully use the weak supervision from image-level labels, we present a novel method called WsMCA to construct multi-level curriculums for revealing diverse properties of target images, as well as adapt the segmentation

TABLE IV: The ablation study on `GTA5` → `Cityscapes` over ResNet-101. The first row displays the "Source-Only" results; the second row displays the results of the domain adversarial training baseline; the third to the fifth rows display the performance contribution of each design for weakly-supervised domain adaptation.

| $\mathcal{L}_{seg}$ | $\mathcal{L}_{adv}$ | $\mathcal{C}_{img}$ | $\mathcal{C}_{reg}$ | $\mathcal{C}_{pix}$ | road | sdwk | bldng | wall | fence | pole | light | sign | vgttn | trrn | sky | person | rider | car | truck | bus | train | mcycl | bcycl | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ✓ | | | | | 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 |
| ✓ | ✓ | | | | 88.0 | 31.6 | 79.6 | **25.5** | 21.6 | 25.4 | 26.0 | 14.3 | 82.6 | 34.1 | 74.0 | 55.3 | 13.9 | 79.7 | 29.4 | 38.3 | 0.9 | 26.2 | 24.9 | 40.6 |
| ✓ | ✓ | ✓ | | | 87.9 | 36.3 | 82.1 | 23.3 | 29.8 | 28.0 | 37.6 | 36.4 | 82.8 | 31.0 | 81.3 | 61.7 | 32.4 | 82.4 | **49.7** | 50.5 | 14.1 | 34.0 | 40.9 | 48.5 |
| ✓ | ✓ | ✓ | ✓ | | 88.6 | 38.7 | **83.0** | 22.0 | **30.3** | 28.8 | 36.4 | 36.7 | 82.5 | 32.2 | 80.7 | 61.8 | 32.7 | 82.1 | 48.0 | **52.8** | 14.4 | **34.2** | 46.5 | 49.1 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **90.3** | **43.4** | 82.4 | 24.1 | 27.1 | **32.9** | **37.8** | **41.0** | **83.2** | 34.1 | 80.8 | **62.0** | **33.7** | **83.3** | 47.7 | 51.3 | 15.0 | 33.8 | **50.3** | **50.2** |

TABLE V: Sensitivity analysis on `GTA5` → `Cityscapes` over ResNet-101. In each row, the corresponding component is progressively incorporated. The sensitivity analysis is conducted through changing the corresponding hyper-parameters, while fixing the others to the values used in the experiments.

| | | | | | |
|---|---|---|---|---|---|
| $\mathcal{C}_{img}$ | $\lambda_{img}$ | 0.05 | 0.1 | 0.2 | 0.4 |
| | mIOU | 47.2 | 48.1 | 48.5 | 47.4 |
| $\mathcal{C}_{reg} + \mathcal{C}_{pix}$ | $\lambda_{reg}(\lambda_{pix})$ | 0.00025 | 0.0005 | 0.001 | 0.002 |
| | mIOU | 49.9 | 50.2 | 50.0 | 49.7 |
| | $\gamma$ | 0.6 | 0.7 | 0.8 | 0.9 |
| | mIOU | 49.3 | 50.2 | 49.7 | 49.4 |
| | Patch size | 50×50 | 100×100 | 150×150 | 200×200 |
| | mIOU | 49.5 | 49.8 | 50.0 | 50.2 |

network towards better pixel-level prediction on target images. Extensive experiments clearly demonstrate that our WsMCA method can effectively leverage weak supervision from target images for cross-domain segmentation.

## Acknowledgments

This work was supported by the Major Project for New Generation of AI (No. 2018AAA0100400) and the National Natural Science Foundation of China (No.11829101, 11931014 and 61772118) and the Fundamental Research Funds for the Central Universities of China (No. JBK1806002).

## References

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017. 1

[2] T. Zhang, G. Lin, J. Cai, T. Shen, C. Shen, and A. C. Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *arXiv preprint arXiv:1803.02563*, 2018. 1, 2

[3] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635-1643. 1, 2

[4] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 549-565. 1, 2

[5] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159-3167. 1, 2

[6] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 102-118. 1

[7] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234-3243. 1, 6

[8] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016. 1, 3

[9] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Unsupervised domain adaptation for semantic segmentation with gans," *arXiv preprint arXiv:1711.06969*, 2017. 1, 3

[10] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6810-6818. 1, 3

[11] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the International Conference on Computer Vision*, 2017, 2039-2049. 1, 2, 3, 4, 7, 8

[12] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 289-305. 1, 2, 3, 4, 5, 6, 7

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440. 2, 6

[14] 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834-848, 2018. 2, 6

[15] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925-1934. 2

[16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. 2

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881-2890. 2

[18] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015. 2

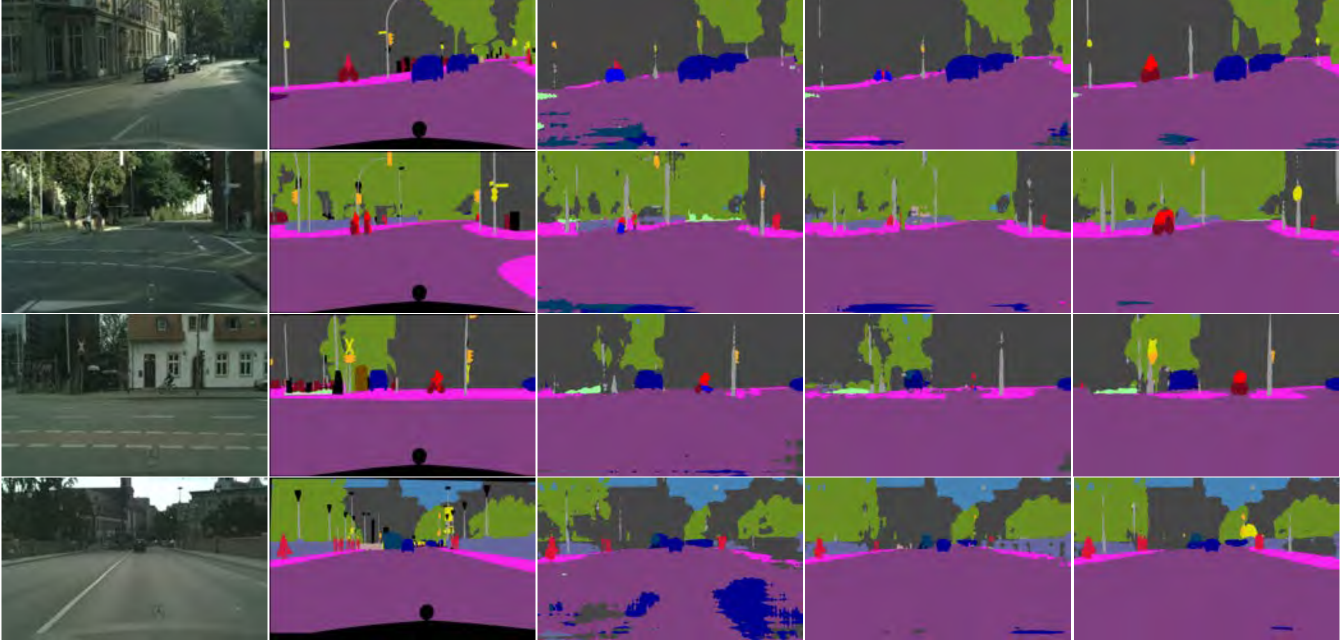
[19] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529-1537. 2

[20] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," *arXiv preprint arXiv:1412.7144*, 2014. 2, 7, 8

[21] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the International Conference on Computer Vision*, 2015, pp. 1742-1750. 2

[22] S. Hong, J. Oh, H. Lee, and B. Han, "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3204-3212. 2

[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921-2929. 2, 7, 8



(a) Target images (b) Ground truth (c) NoAdapt (d) NoWeak (e) Full model

Fig. 5: Qualitative segmentation results on the `GTA5` → `Cityscapes` setting. (a) target images. (b) ground truth. (c) "NoAdapt" predictions of segmentation network trained with only the source domain. (d) segmentation results by removing curriculums constructed with image-level supervision. (e) segmentation results of our full model.

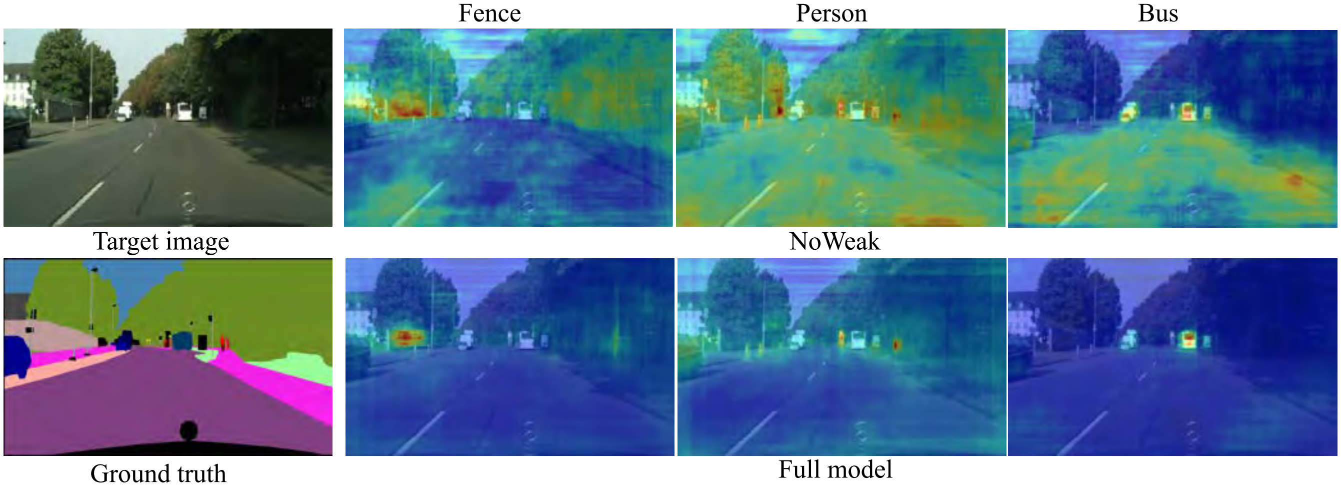


Fig. 6: The activation maps of sample target images under different classes. The activation maps from the top row are obtained by removing curriculums constructed with image-level supervision.

[24] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, "Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. [Online]. Available: https://doi:10.1109/TCSVT.2019.2962073 2

[25] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, and X. Li, "Weakly supervised salient object detection with spatiotemporal cascade neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1973-1984, 2019. 2

[26] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014. 3

[27] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014. 3

[28] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2765-2773. 3

[29] L. Zhang, P. Wang, W. Wei, H. Lu, C. Shen, A. van den Hengel, and Y. Zhang, "Unsupervised domain adaptation using robust classwise matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1339-1349, 2018. 3

[30] W. Deng, L. Zheng, Y. Sun, and J. Jiao, "Rethinking triplet loss for domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. [Online]. Available: https://doi:10.1109/TCSVT.2020.2968484 3

[31] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2507-2516. 3, 6, 7, 8

[32] 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 *Proceedings of the International Conference on Computer Vision*, 2017, pp. 2011-2020. 3, 6

[33] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," *arXiv preprint arXiv:1712.02560*, vol. 3, 2017. 3, 6

[34] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6778-6787. 3, 6, 7, 8

[35] L. Du, J. Tan, H. Yang, J. Feng, X. Xue, Q. Zheng, X. Ye, and X. Zhang, "Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 982-991. 3, 6, 7, 8

[36] 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 preprint arXiv:1802.10349*, 2018. 3, 4, 6, 7, 8

[37] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517-2526.. 3, 6, 7, 8

[38] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4500-4509. 3

[39] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017. 3, 6, 7

[40] Q. Lian, F. Lv, L. Duan, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A nonadversarial approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6758-6767. 3

[41] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M. Sun, "Cross-domain complementary learning using pose for multi-person part segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. [Online]. Available: https://doi:10.1109/TCSVT.2020.2995122 3

[42] K. Lee, G. Ros, J. Li, and A. Gaidon, "SPIGAN: privileged adversarial learning from simulation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rkxoNnC5FQ 3

[43] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "DADA: depth-aware domain adaptation in semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7363-7372. 3, 6, 7, 8

[44] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4376-4386, 2019. 3, 7, 8

[45] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713-1721. 4

[46] Y. Chen, W. Li, and L. Van Gool, "Road: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7892-7901. 6, 7, 8

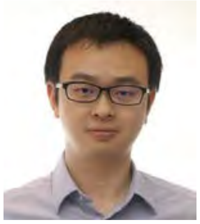
[47] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis, "Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 518-534. 6, 7

[48] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1900-1909. 7, 8

[49] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1456-1465. 6, 7, 8

[50] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213-3223. 6

**Fengmao Lv** received the bachelor's and PhD degrees from the University of Electronic Science and Technology of China, in 2013 and 2018, respectively. He is currently an Assistant Professor at Southwestern University of Finance and Economics. His research interests include computer vision and transfer learning.



**Guosheng Lin** is an Assistant Professor at School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received his PhD degree at The University of Adelaide in 2014. He also received B.Eng. and M.Eng. degrees from the South China University of Technology in 2007 and 2010, respectively. His research interests are in computer vision and machine learning.

**Peng Liu** received the bachelor's degree from the University of Electronic Science and Technology of China in 2018. He is currently a PhD student with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include computer vision and machine learning.

**Guowu Yang** received the PhD degree in computer science from the Portland State University in 2005. He is currently a Full Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include quantum computing and artificial intelligence.

**Sinno Jialin Pan** received the PhD degree in computer science from the Hong Kong University of Science and Technology, in 2010. He is an Associate Professor with the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore. Prior to joining NTU, he was a scientist and lab head of text analytics with the Data Analytics Department, Institute for Infocomm Research, Singapore His research interests include transfer learning and its real-world applications.

**Lixin Duan** received the bachelor's degree from the University of Science and Technology of China in 2008, and the PhD degree from Nanyang Technological University in 2012. He is currently a Full Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include machine learning algorithms (especially in transfer learning and domain adaptation) and their applications in computer vision.