Context-Aware Domain Adaptation in Semantic Segmentation

Jinyu Yang[1], Weizhi An[1], Chaochao Yan[1], Peilin Zhao[2], Junzhou Huang[1]
[1]The University of Texas at Arlington
[2]Tencent AI Lab

Abstract

In this paper, we consider the problem of unsupervised domain adaptation in the semantic segmentation. There are two primary issues in this field, i.e., what and how to transfer domain knowledge across two domains. Existing methods mainly focus on adapting domain-invariant features (what to transfer) through adversarial learning (how to transfer). Context dependency is essential for semantic segmentation, however, its transferability is still not well understood. Furthermore, how to transfer contextual information across two domains remains unexplored. Motivated by this, we propose a cross-attention mechanism based on self-attention to capture context dependencies between two domains and adapt transferable context. To achieve this goal, we design two cross-domain attention modules to adapt context dependencies from both spatial and channel views. Specifically, the spatial attention module captures local feature dependencies between each position in the source and target image. The channel attention module models semantic dependencies between each pair of cross-domain channel maps. To adapt context dependencies, we further selectively aggregate the context information from two domains. The superiority of our method over existing state-of-the-art methods is empirically proved on "GTA5 to Cityscapes" and "SYNTHIA to Cityscapes".

1. Introduction

Semantic segmentation aims to predict pixel-level labels for the given images [22, 3], which has been widely recognized as one of the fundamental tasks in computer vision. Unfortunately, the manual pixel-wise annotation for large-scale segmentation datasets is extremely time-consuming and requires massive amounts of labor efforts. As a tradeoff, synthetic datasets [32, 33] with freely-available labels offer a promising alternative by providing considerable data for model training. However, the domain discrepancy between synthetic (source) and real (target) images is still the central challenge to effectively transfer knowledge across domains. To overcome this limitation, the key idea of existing methods is to leverage knowledge from a source domain to enhance the learning performance of a target domain. Such a strategy is mainly inspired by the recent advances in unsupervised domain adaptation for image classification [31].

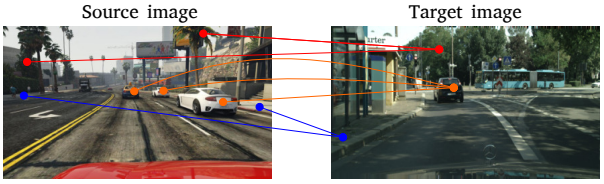

Figure 1. An example of cross-domain context. The source and target images share similar context information at the spatial and semantic level. The red line, orange line, and blue line denote vegetation, car, and sidewalk across two domains, respectively.

Conventional domain adaptation methods in image classification attempt to learn domain-invariant feature representations by directly minimizing the representation distance between two domains [39, 23, 24], encouraging a common feature space through an adversarial objective [11, 38], or automatically determining what and where to transfer via meta-learning [48, 16]. Motivated by this, various domain adaptation methods for semantic segmentation are proposed recently. Among them, the most common practices are based on feature alignment [14, 55], structure adaptation [37, 5], adversarial learning [41, 17, 15, 36], curriculum adaptation [51, 19], self training [56, 18, 46, 30], and image-to-image translation [1, 52, 18, 6, 4, 46]. Despite remarkable performance improvement achieved by these methods, they fail to explicitly consider the contextual dependencies across the source and target domains which is essential for scene understanding [49, 53]. As illustrated in Figure 1, the source and target images share a much similar semantic context such as vegetation, car, and sidewalk, although their appearances (*e.g.*, scale, texture, and illumination) are quite different. However, how to adapt context information across two domains remains unexplored.

Inspired by this, we propose a novel domain adaptation framework named cross-domain attention network (CDANet), designed for urban-scene semantic segmentation. The key idea of CDANet is to leverage cross-domain context

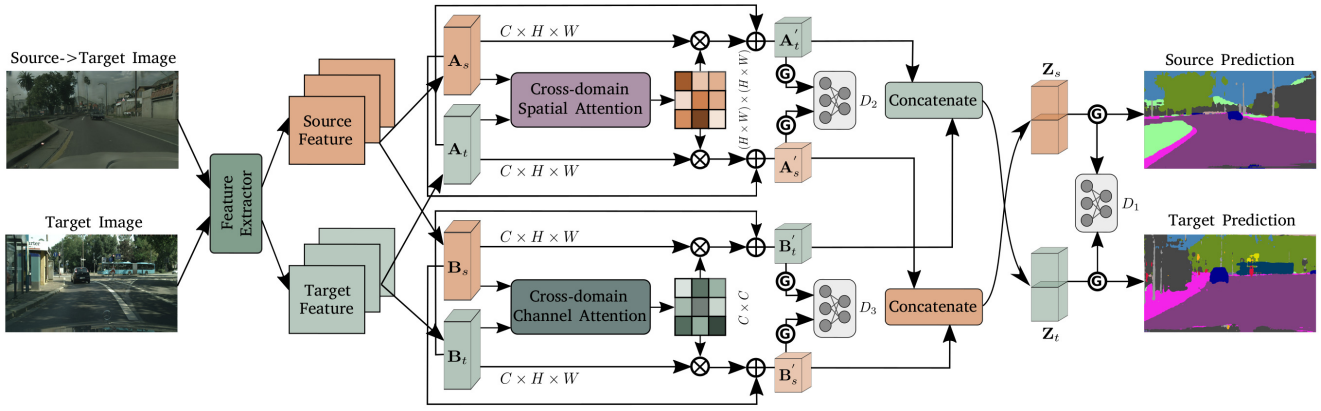

Figure 2. An overview of the proposed framework. It applies a feature extractor (*i.e.*, ResNet101 or VGG16) to learn source and target features. Two cross-domain attention modules (*i.e.*, CD-SAM and CD-CAM) are designed to adapt spatial and semantic context information across source and target domains. A classifier G is used to predict segmentation output based on the features from CD-SAM and CD-CAM. Our framework contains three discriminators (*i.e.*, D_1, D_2, and D_3) for output adaptation by enforcing the source output be indistinguishable from the target output.

dependencies from both a local and global perspective. To achieve this goal, we innovatively design a cross-attention mechanism which contains two cross-domain attention modules to capture mutual context dependencies between source and target domains. Given that same objects with different appearances and scales often share similar features, we introduce a cross-domain spatial attention module (CD-SAM) to capture local feature dependencies between any two positions in a source image and a target image. The CD-SAM involves two directions (*i.e.*, "source-to-target" and "target-to-source") to adaptively aggregate cross-domain features to learn common context information. On the forward direction (or "source-to-target"), CD-SAM updates the feature at each position in the source image as the weighted sum of features at all positions in the target image. The weights are computed based on the similarity of source and target features at each position. Similarly, the backward direction (or "target-to-source") updates the target feature at each position based on the attention to features at all positions in the source image. In consequence, spatial contexts from the source domain are encoded in the target domain, and vice versa. To model the associations between different semantic responses across two domains, we introduce a cross-domain channel attention module (CD-CAM) which has the same bidirectional structure as CD-SAM. The CD-CAM is designed for contextual information aggregation through capturing the channel feature dependencies between any two channel maps in the source and target image. In such a way, common semantic contexts are shared by both domains. CD-SAM and CD-CAM play a complementary role for context adaptation and their outputs are further merged to provide better feature representations for scene understanding.

Our main contributions are summarized as follows: (i) We propose a novel cross-attention mechanism that enables to transfer of context dependencies across two domains. This is the first-of-its-kind study that investigates the transferability of context information in the domain adaptation; (ii) Two cross-domain attention modules are proposed to capture and adapt context dependencies at both spatial and channel levels. This allows us to learn the common semantic context shared by source and target domains; and (iii) Comprehensive empirical studies demonstrate the superiority of our method over the existing state of the art on two benchmark settings, *i.e.*, "GTA5 to Cityscapes" and "SYNTHIA to Cityscapes".

2. Related Work

Domain Adaptation for Semantic Segmentation Inspired by the Generative Adversarial Network [12], Hoffman *et al*. [14] propose the first domain adaptation model for semantic segmentation by learning domain-invariant features through adversarial training. To rule out task-independent factors during feature alignment, SIBAN [25] purifies significance-aware features before the adversarial adaptation to facilitate feature adaptation and stabilize the adversarial training. However, these global adversarial methods ignore to align the category-level joint distribution, which may disturb well-aligned features. To alleviate this problem, Luo *et al*. propose a category-level adversarial network to encourage local semantic consistency through reweighting the adversarial loss for each feature [26]. Similarly, [43] proposes a fine-grained adversarial learning strategy for class-level feature alignment. Based on the hypothesis that structure information plays an essential role in semantic segmentation, Chang *et al*. adapt structure information by learning domain-invariant structure [2]. This is achieved by disentangling the domain-invariant structure of a given image from its domain-specific texture information. AdaptSetNet

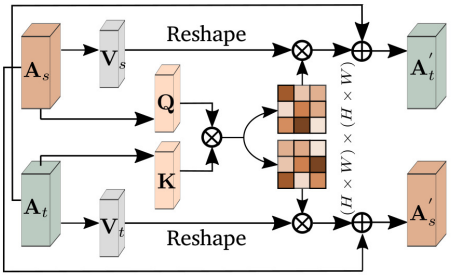

Figure 3. Cross-domain spatial attention module.

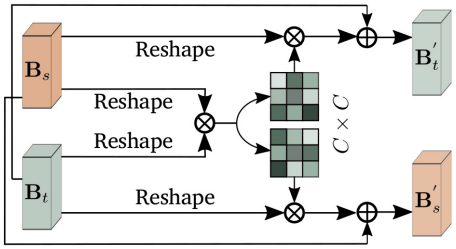

Figure 4. Cross-domain channel attention module.

moves forward by further considering structured output adaptation which is based on the observation that segmentation outputs of the source and target domains share substantial similarities [37]. Different from AdaptSetNet, we apply three domain discriminators to perform output adaptation on the segmentation outputs from CD-SAM, CD-CAM, and the aggregation of these two modules.

Most recently, image-to-image translation [54] has proved its effectiveness in domain adaptation [1, 45, 6]. The key idea is to translate images from the source domain to the target domain by using an image translation model and use the translated images for adapting cross-domain knowledge through a segmentation adaptation model. Rather than keeping the image translation model unchanged after obtaining translated images, BDL [18] applies a bidirectional learning framework to alternatively optimize the image translation model and the segmentation model. Similar to [56, 30], a self-supervised learning strategy is also used in BDL to generate pseudo labels for target images and re-training the segmentation model with these labels. Although BDL achieves the new state of the art, it is limited in its ability to consider the cross-domain context dependencies. To overcome this limitation, we introduce two cross-domain attention modules to adapt context information between source and target domains.

Context-Aware Embedding It has been long known that context information plays an important role in perceptual tasks such as semantic segmentation [29]. Zhang *et al*. [49] propose a context encoding module to capture the semantic context of scenes and selectively emphasize or de-emphasize class-dependent feature maps. To aggregate image-adapted context, MSCI [20] further considers multi-scale context embedding and spatial relationships among super-pixels in a given image. Following the success of attention mechanism [40] in image generation [50] and sentence embedding [21], recent studies have highlighted the potential of self-attention in capturing context dependencies [10, 53]. Specifically, Zhao *et al*. [53] introduce a point-wise spatial attention network to aggregate long-range contextual information. Their model mainly draws its strength from the self-adaptively predicted attention maps which can take full advantage of both nearby and distant information of each pixel. DANet [10] adaptively integrates local features with their global dependencies through a position attention module and a channel attention module. These two modules are considered to be able to capture spatial and semantic interdependencies, and in turn, facilitate scene understanding. Similarly, CBAM [44] sequentially infers attention maps along the channel and spatial dimensions in order to adaptively refine the intermediate features. As opposed to capturing contextual information within a single domain as previously reported, we design an innovative cross-attention mechanism to model context dependencies between two different domains, which is essential for context adaptation.

3. Methodology

In this section, we begin by briefing the key idea of our framework. We then detail the proposed cross-attention mechanism which contains two cross-domain attention modules for adapting context dependencies between a source and a target domain.

3.1. Overview

Given a set of source-domain images $\mathcal{X}_s$ with pixel-wise labels Y_s and a set of target-domain images $\mathcal{X}_t$ without any annotation. Our goal is to train a segmentation model that can provide accurate prediction to $\mathcal{X}_t$. To achieve this, $\mathcal{X}_s$ is first translated from the source domain to the target domain using CycleGAN [54]. The translated images $\mathcal{X}_s^{'} = \mathcal{F}(\mathcal{X}_s)$ (where $\mathcal{F}$ denotes the image translation model) share the same semantic labels with $\mathcal{X}_s$ but with common visual appearance as $\mathcal{X}_t$. Motivated by the self-training strategy, we follow the same idea in [18, 30] to generate pseudo labels Y_t^{st} for $\mathcal{X}_t$ with high prediction confidence. Coordinated with these translated images and pseudo labels, we introduce a cross-attention mechanism for domain adaptation of semantic segmentation by leveraging cross-domain contextual information (Figure 2). First, a feature extractor E is applied to get source feature $E(\mathcal{X}_s^{'})$ and target feature $E(\mathcal{X}_t)$ which are 1/8 of the corresponding input image size. Then a linear interpolation is applied to $E(\mathcal{X}_s^{'})$ and $E(\mathcal{X}_t)$ to match their spatial size. After that, two parallel convolution layers are applied to $E(\mathcal{X}_s^{'})$ and $E(\mathcal{X}_s)$ to generate feature pairs $\{\mathbf{A}_s, \mathbf{A}_t\}$ and $\{\mathbf{B}_s, \mathbf{B}_t\}$, respectively. $\{\mathbf{A}_s, \mathbf{A}_t\}$ is then fed

Table 1. The performance comparison by adapting from GTA5 to Cityscapes. Two base architectures (i.e., VGG16 and ResNet101) are used in our study. The comparison is performed on 19 common classes between source and target domains. We use per-class IoU and mean IoU (mIoU) for the performance measurement. The best result in each column is highlighted in bold.

GTA5 to Cityscapes																					
	Architecture	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	**mIoU**
FCNs wild [14]	VGG16	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
CDA [51]		74.9	22.0	71.4	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
AdaptSegNet [37]		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 [1]		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
LSD [34]		88.0	30.5	78.6	25.2	23.5	16.7	23.5	11.6	78.7	27.2	71.9	51.3	19.5	80.4	19.8	18.3	0.9	20.8	18.4	37.1
PyCDA [19]		86.7	24.8	80.9	21.4	27.3	**30.2**	26.6	21.1	**86.6**	28.9	58.8	53.2	17.9	80.4	18.8	22.4	4.1	9.7	6.2	37.2
CrDoCo [6]		89.1	33.2	80.1	26.9	25.0	18.3	23.4	12.8	77.0	29.1	72.4	55.1	20.2	79.9	22.3	19.5	1.0	20.1	18.7	38.1
BDL [18]		89.2	40.9	81.2	29.1	19.2	14.2	29.0	19.6	83.7	35.9	80.7	54.7	23.3	82.7	**25.8**	28.0	2.3	**25.7**	19.9	41.3
FDA [47]		86.1	35.1	80.6	30.8	20.4	27.5	30.0	**26.0**	82.1	30.3	73.6	52.5	21.7	81.7	24.0	30.5	**29.9**	14.6	24.0	42.2
FADA [43]		**92.3**	**51.1**	**83.7**	33.1	**29.1**	28.5	28.0	21.0	82.6	32.6	**85.3**	55.2	**28.8**	**83.5**	24.4	**37.4**	0.0	21.1	15.2	43.8
Ours		90.1	46.7	82.7	**34.2**	25.3	21.3	**33.0**	22.0	84.4	**41.4**	78.9	**55.5**	25.8	83.1	24.9	31.4	20.6	25.2	**27.8**	**44.9**
AdaptSegNet [37]	ResNet101	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
CLAN [26]		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
IntraDA [30]		90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	**85.7**	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
MaxSquare [28]		89.4	43.0	82.1	30.5	21.3	30.3	34.7	24.0	85.3	39.4	78.2	63.0	22.9	84.6	**36.4**	43.0	5.5	34.7	33.5	46.4
BDL [18]		91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	**43.6**	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
FADA [43]		**92.5**	47.5	**85.1**	**37.6**	**32.8**	33.4	33.8	18.4	85.3	37.7	**83.5**	**63.2**	**39.7**	**87.5**	32.9	47.8	1.6	**34.9**	39.5	49.2
FDA [47]		**92.5**	**53.3**	82.4	26.5	27.6	**36.4**	**40.6**	**38.9**	82.3	39.8	78.0	62.6	34.4	84.9	34.1	**53.1**	**16.9**	27.7	**46.4**	**50.4**
Ours		91.3	46.0	84.5	34.4	29.7	32.6	35.8	36.4	84.5	43.2	83.0	60.0	32.2	83.2	35.0	46.7	0.0	33.7	42.2	49.2

into CD-SAM to adapt spatial-level context, while CD-CAM adapts channel-level context based on $\{\mathbf{B}_s, \mathbf{B}_t\}$.

For each module, two directions, *i.e.*, forward direction ("source-to-target") and backward direction ("target-to-source") are involved. Take the CD-SAM as an example, an energy map is first obtained based on $\{\mathbf{A}_s, \mathbf{A}_t\}$. This energy map is further divided into two attention matrices denoted by $\Gamma_{s\to t}$ and $\Gamma_{t\to s}$. During the forward direction, we perform a matrix multiplication between target features and $\Gamma_{s\to t}$. The result is then summed with the original source features in an element-wise manner. For the backward direction, a matrix multiplication is conducted between source features and $\Gamma_{t\to s}$. After that, an element-wise summation between the obtained results and original target features is carried out. The CD-CAM follows the same setting above except that the energy map is calculated in the channel dimension. The final source feature and target feature are obtained by aggregating the outputs from these two attention modules, which are then fed into a classifier G for semantic segmentation.

3.2. Cross-Domain Spatial Attention Module

The goal of CD-SAM is to adapt spatial contextual information across two domains. To achieve this, we introduce the forward direction ("source-to-target") to augment source features by selectively aggregating target features based on their similarities. We further introduce the backward direction ("target-to-source") to update target features by aggregating source features in the same way.

The architecture of CD-SAM is illustrated in Figure 3. Given $\mathbf{A}_s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{A}_t \in \mathbb{R}^{C\times H\times W}$ (C denotes the channel number and $H \times W$ indicates the spatial size), two parallel convolution layers are applied to generate $\mathbf{Q} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{K} \in \mathbb{R}^{C\times H\times W}$, respectively. $\mathbf{A}_s$ and $\mathbf{A}_t$ are also fed into another convolution layer to obtain $\mathbf{V}_s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{V}_t \in \mathbb{R}^{C\times H\times W}$. We reshape $\mathbf{Q}$, $\mathbf{V}_s$, $\mathbf{K}$, and $\mathbf{V}_t$ to $C \times N$, where $N = H \times W$. To determine spatial context relationships between each position in $\mathbf{A}_s$ and $\mathbf{A}_t$, an energy map $\Phi \in \mathbb{R}^{N\times N}$ is formulated as $\Phi = \mathbf{Q}^T\mathbf{K}$, where $\Phi^{(i,j)}$ measure the similarity between i^{th} position in $\mathbf{A}_s$ and j^{th} position in $\mathbf{A}_t$. To augment $\mathbf{A}_s$ with spatial context information from $\mathbf{A}_t$ and vice versa, a bidirectional feature adaptation is defined as follows.

During the forward direction, we first define the "source-to-target" spatial attention map as,

$$\Gamma_{s\to t}^{(i,j)} = \frac{exp(\Phi^{(i,j)})}{\sum_{j=1}^{N_t} exp(\Phi^{(i,j)})}, \tag{1}$$

where $\Gamma_{s\to t}^{(i,j)}$ indicates the impact of i^{th} position in $\mathbf{A}_s$ to j^{th} position in $\mathbf{A}_t$. To capture spatial context in the target domain, we update $\mathbf{A}_s$ as,

$$\mathbf{A}_s^{'} = \mathbf{A}_s + \lambda_s \mathbf{V}_t \Gamma_{s\to t}^{T}, \tag{2}$$

where λ_s leverages the importance of target-domain context and original source features. In this regime, each position in $\mathbf{A}_s^{'}$ has a global context view of target features.

Table 2. The performance comparison by adapting from SYNTHIA to Cityscapes. Two base architectures (i.e., VGG16 and ResNet101) are used in our study. The comparison is performed on 16 common classes for VGG16 and 13 common classes for ResNet101.

SYNTHIA to Cityscapes																	
	Architecture	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	sky	person	rider	car	bus	motorbike	bicycle	**mIoU**
DCAN [45]	VGG16	79.9	30.4	70.8	1.6	**0.6**	22.3	6.7	23.0	76.9	73.9	41.9	16.7	61.7	11.5	10.3	38.6	35.4
PyCDA [19]		80.6	26.6	74.5	2.0	0.1	18.1	**13.7**	14.2	80.8	71.0	48.0	19.0	72.3	22.5	12.1	18.1	35.9
DADA [42]		71.1	29.8	71.4	3.7	0.3	**33.2**	6.4	15.6	81.2	78.9	52.7	13.1	75.9	25.5	10.0	20.5	36.8
GIO-Ada [4]		78.3	29.2	76.9	**11.4**	0.3	26.5	10.8	17.2	81.7	81.9	45.8	15.4	68.0	15.9	7.5	30.4	37.3
TGCF-DA [7]		**90.1**	**48.6**	80.7	2.2	0.2	27.2	3.2	14.3	**82.1**	78.4	54.4	16.4	**82.5**	12.3	1.7	21.8	38.5
BDL [18]		72.0	30.3	74.5	0.1	0.3	24.6	10.2	25.2	80.5	80.0	54.7	23.2	72.7	24.0	7.5	44.9	39.0
FADA [43]		80.4	35.9	**80.9**	2.5	0.3	30.4	7.9	22.3	81.8	**83.6**	48.9	16.8	77.7	**31.1**	13.5	17.9	39.5
FDA [47]		84.2	35.1	78.0	6.1	0.4	27.0	8.5	22.1	77.2	79.6	55.5	19.9	74.8	24.9	**14.3**	40.7	40.5
Ours		73.0	31.1	77.1	0.2	0.5	27.0	11.3	**27.4**	81.2	81.0	**59.0**	**25.6**	75.0	26.3	10.1	**47.4**	**40.8**
SIBAN [25]	ResNet101	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 [26]		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
MaxSquare [28]		82.9	40.7	80.3	✗	✗	✗	12.8	18.2	82.5	82.2	53.1	18.0	79.0	31.4	10.4	35.6	48.2
IntraDA [30]		84.3	37.7	79.5	✗	✗	✗	9.2	8.4	80.0	**84.1**	57.2	23.0	78.0	38.1	20.3	36.5	48.9
DADA [42]		**89.2**	44.8	81.4	✗	✗	✗	8.6	11.1	81.8	84.0	54.7	19.3	79.7	40.7	14.0	38.8	49.8
BDL [18]		86.0	**46.7**	80.3	✗	✗	✗	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	51.4
FDA [47]		79.3	35.0	73.2	✗	✗	✗	19.9	24.0	61.7	82.6	**61.4**	31.1	83.9	40.8	**38.4**	**51.1**	**52.5**
FADA [43]		84.5	40.1	**83.1**	✗	✗	✗	**20.1**	**27.2**	**84.8**	84.0	53.5	22.6	**85.4**	**43.7**	26.8	27.8	**52.5**
Ours		82.5	42.2	81.3	✗	✗	✗	18.3	15.9	80.6	83.5	**61.4**	**33.2**	72.9	39.3	26.6	43.9	52.4

For the backward direction, the "target-to-source" spatial attention map is formulated as,

$$\Gamma_{t\to s}^{(i,j)} = \frac{exp(\Phi^{(i,j)})}{\sum_{i=1}^{N_s} exp(\Phi^{(i,j)})}, \tag{3}$$

where $\Gamma_{t\to s}^{(i,j)}$ indicates to what extent the j^{th} position in $\mathbf{A}_t$ attends to the i^{th} position in $\mathbf{A}_s$. Similarly, $\mathbf{A}_t$ is updated by,

$$\mathbf{A}_t^{'} = \mathbf{A}_t + \lambda_t \mathbf{V}_s \Gamma_{t\to s}, \tag{4}$$

where λ_t leverages the importance of source-domain context and original target features. As a consequence, each position in $\mathbf{A}_s^{'}$ and $\mathbf{A}_t^{'}$ is a combination of their original feature and the weighed sum of features from the opposite domain. Therefore, $\mathbf{A}_s^{'}$ and $\mathbf{A}_t^{'}$ allow us to encode the spatial context of both source and target domains.

3.3. Cross-Domain Channel Attention Module

Given $\mathbf{B}_s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{B}_t \in \mathbb{R}^{C\times H\times W}$, the CD-CAM is designed to adapt semantic context between source and target domains (Figure 4) by following the same bidirectional structure as CD-SAM. Different from CD-SAM that applies convolution layers to obtain $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}_s$, and $\mathbf{V}_t$ before measuring spatial relationships. Here, $\mathbf{B}_s$ and $\mathbf{B}_t$ are directly used to capture their semantical context relationships, which allows us to maintain interdependencies between channel maps [10]. Specifically, we reshape both $\mathbf{B}_s$ and $\mathbf{B}_t$ to $C \times N$, where $N = H \times W$. The energy map is defined as $\Theta = \mathbf{B}_t \mathbf{B}_s^T \in \mathbb{R}^{C\times C}$, where $\Theta^{(i,j)}$ denotes the similarity between i^{th} channel in $\mathbf{B}_s$ and j^{th} channel in $\mathbf{B}_t$.

For the forward direction, the "source-to-target" attention map is given by,

$$\Psi_{s\to t}^{(i,j)} = \frac{exp(\Theta^{(i,j)})}{\sum_{j=1}^{C} exp(\Theta^{(i,j)})}, \tag{5}$$

where $\Psi_{s\to t}^{(i,j)}$ measures the impact of i^{th} channel in $\mathbf{B}_s$ to j^{th} channel in $\mathbf{B}_t$. To model the cross-domain semantic context dependencies, $\mathbf{B}_s$ is updated by,

$$\mathbf{B}_s^{'} = \mathbf{B}_s + \xi_s \Psi_{s\to t} \mathbf{B}_t, \tag{6}$$

where ξ_s leverages the associations between target-domain semantic information and original source features. As a consequence, each channel in $\mathbf{B}_s^{'}$ is augmented by selectively aggregating semantic information from $\mathbf{B}_t$.

During the backward direction, the "target-to-source" attention map is,

$$\Psi_{t\to s}^{(i,j)} = \frac{exp(\Theta^{(i,j)})}{\sum_{i=1}^{C} exp(\Theta^{(i,j)})} \tag{7}$$

To take semantic context in $\mathbf{B}_s$ into consideration, we have

$$\mathbf{B}_t^{'} = \mathbf{B}_t + \xi_t \Psi_{t\to s}^T \mathbf{B}_s, \tag{8}$$

where ξ_t leverages the associations between original target features and semantic contexts from the source domain. It is noteworthy that by considering cross-domain semantic context, our framework is able to further reduce domain discrepancy from the context perspective.

Table 3. Ablation study on "GTA5 to Cityscapes".

GTA5 to Cityscapes			
Base	CD-SAM	CD-CAM	**mIoU**
VGG16			41.3
	✓		43.7
		✓	43.6
	✓	✓	44.9
ResNet101			48.5
	✓		49.0
		✓	48.8
	✓	✓	49.2

Table 4. Ablation study on "SYNTHIA to Cityscapes".

SYNTHIA to Cityscapes			
Base	CD-SAM	CD-CAM	**mIoU**
VGG16			39.0
	✓		40.2
		✓	40.0
	✓	✓	40.8
ResNet101			51.4
	✓		51.8
		✓	52.0
	✓	✓	52.4

3.4. Aggregation of Spatial and Channel Context

To take full advantage of spatial and channel context information, we aggregate the outputs from these two cross-domain attention modules. Specifically, $\mathbf{A}_s^{'}$ and $\mathbf{B}_s^{'}$ are concatenated and then fed into a convolution layer to generate the enhanced source feature $\mathbf{Z}_s \in \mathbb{R}^{C\times H\times W}$. Obviously, $\mathbf{Z}_s$ is enriched by spatial and semantic context dependencies from both source and target domains. The same operation is also performed on $\mathbf{A}_t^{'}$ and $\mathbf{B}_t^{'}$ to obtain $\mathbf{Z}_t \in \mathbb{R}^{C\times H\times W}$.

3.5. Training Objective

Our framework contains a segmentation loss $\mathcal{L}_{seg}$ and an adversarial loss $\mathcal{L}_{adv}$. We first feed $\mathbf{Z}_s$ and $\mathbf{Z}_t$ into the classifier G to predict their segmentation outputs $G(\mathbf{Z}_s)$ and $G(\mathbf{Z}_t)$. The segmentation loss of $G(\mathbf{Z}_s)$ is defined as:

$$\mathcal{L}_{seg}(G(\mathbf{Z}_s), Y_s) = -\sum_{i=1}^{H\times W}\sum_{j=1}^{L} Y_s^{(i,j)} G(\mathbf{Z}_s)^{(i,j)}, \quad (9)$$

where L is the number of label classes. $\mathcal{L}_{seg}(G(\mathbf{Z}_t), Y_s^{st})$ is defined in a similar way. To adapt structured output space [37], a discriminator D_1 is applied to $G(\mathbf{Z}_s)$ and $G(\mathbf{Z}_t)$ to make them be indistinguishable from each other. To achieve this, an adversarial loss $\mathcal{L}_{adv}(G(\mathbf{Z}_s), G(\mathbf{Z}_t))$ is formulated as,

$$\begin{aligned}\mathcal{L}_{adv}(G(\mathbf{Z}_s), G(\mathbf{Z}_t), D_1) = \ & \mathbb{E}[logD_1(G(\mathbf{Z}_s))]+ \\ & \mathbb{E}[log(1-D_1(G(\mathbf{Z}_t)))]\end{aligned} \quad (10)$$

To encourage $\mathbf{A}_s^{'}$, $\mathbf{A}_t^{'}$, $\mathbf{B}_s^{'}$ and $\mathbf{B}_t^{'}$ to encode useful information for semantic segmentation, they are also fed into the classifier G to predict their segmentation outputs. The overall segmentation loss is given by,

$$\begin{aligned}\mathcal{L}_{seg} = & \mathcal{L}_{seg}(G(\mathbf{Z}_s), Y_s) + \mathcal{L}_{seg}(G(\mathbf{Z}_t), Y_t^{st})+ \\ & \mathcal{L}_{seg}(G(\mathbf{A}_s^{'}), Y_s) + \mathcal{L}_{seg}(G(\mathbf{A}_t^{'}), Y_t^{st})+ \\ & \mathcal{L}_{seg}(G(\mathbf{B}_s^{'}), Y_s) + \mathcal{L}_{seg}(G(\mathbf{B}_t^{'}), Y_t^{st})\end{aligned} \quad (11)$$

We also encourage $G(\mathbf{A}_s^{'})$ and $G(\mathbf{A}_t^{'})$ to have similar structured layout, and enforce $G(\mathbf{B}_s^{'})$ to be indistinguishable from $G(\mathbf{B}_t^{'})$. Therefore, the overall adversarial loss can be written as,

$$\begin{aligned}\mathcal{L}_{adv} = & \mathcal{L}_{adv}(G(\mathbf{Z}_s), G(\mathbf{Z}_t), D_1)+ \\ & \mathcal{L}_{adv}(G(\mathbf{A}_s^{'}), G(\mathbf{A}_t^{'}), D_2)+ \\ & \mathcal{L}_{adv}(G(\mathbf{B}_s^{'}), G(\mathbf{B}_t^{'}), D_3),\end{aligned} \quad (12)$$

where D_2 and D_3 are two discriminators. Specifically, D_2 aims to discriminate between $G(\mathbf{A}_s^{'})$ and $G(\mathbf{A}_t^{'})$, while D_3 attempts to distinguish between $G(\mathbf{B}_s^{'})$ and $G(\mathbf{B}_t^{'})$.

Taken them together, the training objective of our framework is:

$$\min_{E,G}\max_{D_1,D_2,D_3} \mathcal{L}_{seg} + \lambda\mathcal{L}_{adv} \quad (13)$$

where λ controls the importance of $\mathcal{L}_{seg}$ and $\mathcal{L}_{adv}$.

4. Experiments

In this section, we evaluate our method on synthetic-to-real domain adaptation for urban scene understanding problem. Extensive empirical experiments and ablation studies are performed to demonstrate out method's superiority over existing state-of-the-art models. We also visualize the cross-domain attention maps to reveal context dependencies between source and target domains.

4.1. Datasets

Two synthetic datasets, *i.e.*, GTA5 [32] and SYNTHIA-RAND-CITYSCAPES [33] are used as the source domain in our study, while the Cityscapes [8] is served as the target domain. Specifically, the GTA5 is collected from a photorealistic open-world game known as Grand Theft Auto V, which contains 24,966 images with pixel-accurate semantic labels. The resolution of each image is 1914 × 1052. SYNTHIA-RAND-CITYSCAPES contains 9,400 images (1280 × 760) with precise pixel-level semantic annotations, which are generated from a virtual city. Cityscapes is a large-scale street scene datasets collected from 50 cities, including 5,000 images with high-quality pixel-level annotations. These images are split into training (2,975 images), validation (500 images), and test (1,525 images) set, each of which with the resolution of 2048 × 1024. Following the same setting as previous studies, only the training set from Cityscapes is

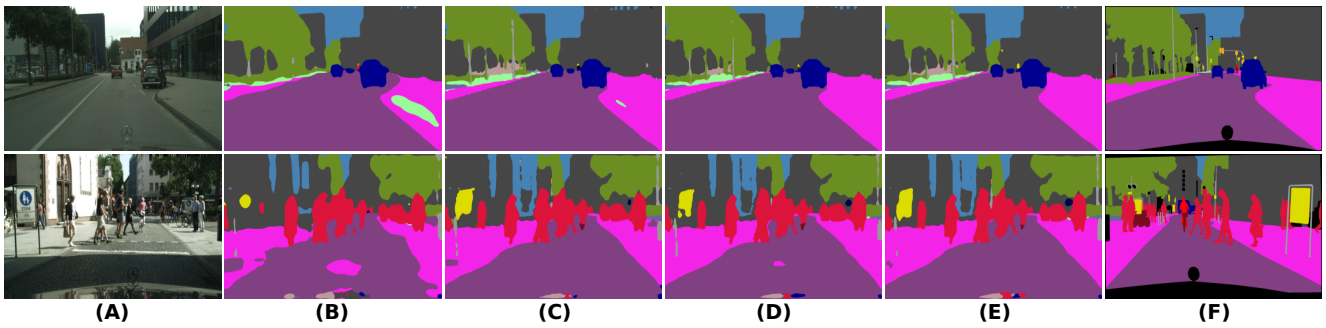

Figure 5. Qualitative comparison between our method and the baseline model BDL [18]. For each given image (A), we present its segmentation output from (B) BDL, (C) our method incorporating CD-SAM only, (D) our method incorporating CD-CAM only, (E) our method considering both CD-SAM and CD-CAM, and the ground truth (F).

used as the target domain, and the validation set is used for performance evaluation.

4.2. Implementation Details

Network Architecture The same CycleGAN architecture [54] as reported in BDL [18] is used to translate images from the source domain to the target domain. DeepLab-VGG16 and DeepLab-ResNet101, which are pre-trained on ImageNet [9], are used as our segmentation network by following the same setting in [37]. Both of them use DeepLab-v2 [3] as classifier, while DeepLab-VGG16 uses VGG16 [35] and DeepLab-ResNet101 uses ResNet101 [13] as the feature extractor. The three discriminators used for structured output adaptation have the identical architecture, each of which has 5 convolution layers with kernel 4×4 and stride of 2. The channel number of each layer is $\{64, 128, 256, 512, 1\}$. Each layer is followed by a leaky ReLU [27] parameterized by 0.2 except the last one. The CD-SAM contains 3 convolution layers with kernel 1×1 and stride of 1 to obtain the query and key-value pairs. The channel number of these convolution layers are $\{128, 128, 1024\}$ and $\{256, 256, 2048\}$ for DeepLab-VGG16 and DeepLab-ResNet101, respectively.

Network Training To train the CycleGAN network, we follow the same setting in BDL [18]. DeepLab-VGG16 is trained using Adam optimizer with initial learning rate 1e-5 and momentum (0.9, 0.99). We apply step decay to the learning rate with step size 50000 and drop factor 0.1. Both DeepLab-ResNet101 and CD-SAM use Stochastic Gradient Descent (SGD) optimizer with momentum 0.9 and weight decay 5e-4. The initial learning rate for DeepLab-ResNet101 and CD-SAM are 2.5e-4 and 1e-4, respectively, and are decreased by the same polynomial policy with power 0.9. For the discriminator, we use an Adam optimizer with momentum (0.9, 0.99). Its initial learning rate is set to 1e-6 for DeepLab-VGG16 and 1e-4 for DeepLab-ResNet101, respectively. We set λ to 0.0001 and 0.001 for DeepLab-VGG16 and DeepLab-ResNet101, respectively.

4.3. Performance Comparison

GTA5 to Cityscapes Our method is first evaluated by using GTA5 as the source domain and Cityscapes as the target domain. The performance is assessed on 19 common classes between these two datasets by following the same evaluation criterion in previous studies [18, 6]. Our method is compared with existing state-of-the-art models by using VGG16 and ResNet101 as the base architectures. As shown in Table 1, our method competes favorably against other models. Specifically, we surpass the mean intersection-over-union (mIoU) of feature alignment-based [14, 34, 26] and curriculum-based methods [51] by a large margin. This observation indicates that simply aligning feature space and label distribution cannot fully transfer domain knowledge in semantic segmentation. Compared to the models [1, 6, 18] that are based on image-to-image translation, our method gains up to 9.5% improvement by using VGG16, revealing that domain discrepancy can be further reduced by considering context adaptation. Similar to [37, 18], we also adapt structured output space in our model, but our method achieves significant performance improvement. This observation reveals the important role of context adaptation in knowledge transfer. It is noteworthy that the prediction of the "train" class is extremely challenging, owing to the limited "train" samples in the source domain. Our method enables to alleviate this limitation by adapting cross-domain context information. Compared to the CyCADA [1], we achieve 16.1% improvement on the "train" class.

SYNTHIA to Cityscapes The superiority of our method is further proved on "SYNTHIA to Cityscapes". It is noteworthy that domain adaptation on "SYNTHIA to Cityscapes" is more challenging than "GTA5 to Cityscapes", owing to the large domain gap between these two domains. Following [18], we consider the 16 and 13 common classes for VGG16 and ResNet101-based models, respectively. As summarized in Table 2, we achieve a performance improvement of 1.8% and 1.0% over BDL [18] with VGG16 and ResNet101 base architectures. One of the most significant difference between

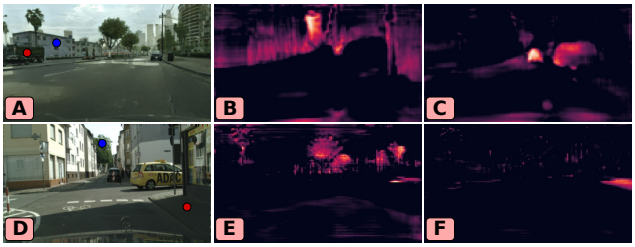

Figure 6. An example of the spatial attention map. Given a source image (A) and a target image (D), we present the source-to-target attention maps (B) and (C) for the blue and red point in (A), respectively. Similarly, we present the target-to-source attention maps (E) and (F) of the blue and red point in (D), respectively.

these two domains is that SYNTHIA has much more 'person' instances than Cityscapes, which makes it hard to transfer common knowledge of the class 'person' by simply aligning marginal distribution or structured output space [18]. In contrast, by considering context information explicitly, we bring 7.3% improvement compared to BDL on this class with ResNet101-based model. This result demonstrates the benefit of explicitly adapting cross-domain context dependencies in semantic segmentation, especially for two domains with significant differences.

4.4. Ablation Study

GTA5 to Cityscapes By incorporating CD-SAM and CD-CAM individually, we get 2.4% and 2.3% performance boost over the VGG16-based baseline (Table 3). Taken them together, the mIoU is further improved to 44.9 mIoU. Similarly, 0.5% and 0.3% improvement is also observed in the ResNet101-based model by considering CD-SAM and CD-CAM. We achieve 49.2 mIoU by integrating both attention modules. To qualitatively demonstrate the superiority of our method, we showcase the examples of its segmentation outputs at different stages in Figure 5. As shown in the figure, our method enables to predict more consistent segmentation outputs than the baseline model and becomes increasingly accurate by incorporating two cross-domain attention modules.

SYNTHIA to Cityscapes For VGG16-based model, CD-SAM and CD-CAM contribute to 1.2% and 1.0% improvement compared to the baseline (Table 4). Our method gains 1.8% improvement by combining them. By applying CD-SAM and CD-CAM to ResNet101, we achieve 51.8 and 52.0 mIoU with 0.4% and 0.6% improvement over the baseline, respectively. It is further boosted to 52.4 mIoU when both of them are considered. Our results reveal that the proposed cross-attention mechanism significantly contributes to domain adaptation in semantic segmentation by adapting context dependencies. Furthermore, the two cross-domain attention modules play a complementary role in capturing context information.

Table 5. Ablation study of λ_s, λ_t, ξ_s, and ξ_t.

$\lambda_s/\lambda_t/\xi_s/\xi_t$	0.1	1	10
mIoU	43.7	44.9	40.6

Visualization of the Cross Attention To fully understand the cross-attention mechanism in our model, we visualize the spatial attention maps in this section. As shown in Figure 6, two images are randomly selected from the source and target domain. Recall that each position in the source feature has a spatial attention map corresponding to all positions in the target feature, and vice versa. We, therefore, select two positions in the source image and visualize their "source-to-target" attention map. For the blue point that is marked on a building in the source image (Figure 6 A), its spatial attention map (Figure 6 B) mainly corresponds to the building in the target image (Figure 6 D). For the red point that is marked on a truck in Figure 6 A, its spatial attention map (Figure 6 C) highlights the cars in Figure 6 D. Similarly, we select another two positions in the target image and conduct the visualization of the "target-to-source" attention map. For the blue point in the target image (Figure 6 D), its attention map (Figure 6 E) focuses on the vegetation in the source image (Figure 6 A). These visualizations demonstrate the power of our method in capturing cross-domain spatial context information.

Parameter Sensitivity Analysis In this section, we perform a sensitivity analysis of λ_s, λ_t, ξ_s, and ξ_t as shown in Table 5. We investigate three different choices, *i.e.*, 0.1, 1, and 10, indicating how much attention should pay for the context information from the opposite domain. Our results reveal that $\lambda_s = \lambda_t = \xi_s = \xi_t = 1$ performs best. The reason is that a small value fails to capture cross-domain context dependencies, while a large value may disturb the original feature. In addition, by setting $\lambda_s = \lambda_t = 0.1$, $\xi_s = \xi_t = 1$, we have mIoU 43.2. We also evaluate the scenario where λ_s, λ_t, ξ_s, and ξ_t are learnable hyperparameters, which gives rise to mIoU 44.0.

5. Conclusion

In this paper, we propose an innovative cross-attention mechanism for domain adaptation by adapting the semantic context. Specifically, we introduce two cross-domain attention modules to capture spatial and channel context between source and target domains. The obtained contextual dependencies, which are shared across two domains, are further adapted to decrease the domain discrepancy. Empirical studies demonstrate that our method achieves the new state-of-the-art performance on "GTA5-to-Cityscapes" and "SYNTHIA-to-Cityscapes".

Acknowledgments This work was partially supported by US National Science Foundation IIS-1718853, the CAREER grant IIS-1553687 and Cancer Prevention and Research Institute of Texas (CPRIT) award (RP190107).

References

[1] Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

[2] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1909, 2019.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 40(4):834–848, 2018.

[4] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1841–1850, 2019.

[5] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1791–1800, 2019.

[7] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pages 2672–2680, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[15] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *International Conference on Machine Learning (ICML)*, 2019.

[17] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. *International Conference on Learning Representations (ICLR)*, 2019.

[18] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6936–6945, 2019.

[19] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.

[20] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018.

[21] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations (ICLR)*, 2017.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015.

[24] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning (ICML)*, 2017.

[25] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, October 2019.

[26] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2507–2516, 2019.

[27] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, 2013.

[28] Hongyang Xue Minghao Chen and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014.

[30] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 2009.

[32] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2016.

[33] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016.

[34] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3752–3761, 2018.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[37] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176, 2017.

[39] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NIPS)*, pages 5998–6008, 2017.

[41] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.

[43] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[45] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[46] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 2020.

[47] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] Wei Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. Transfer learning via learning to transfer. In *International Conference on Machine Learning (ICML)*, pages 5072–5081, 2018.

[49] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, 2018.

[50] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018.

[51] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[52] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[53] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise

spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018.

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2223–2232, 2017.

[55] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[56] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 297–313, 2018.