Contextual-Relation Consistent Domain Adaptation for Semantic Segmentation

Jiaxing Huang[1][0000−0002−8681−0471], Shijian Lu[1][0000−0002−6766−2506], Dayan Guan[1][0000−0001−9752−1520], and Xiaobing Zhang[2][0000−0002−8149−1424]

[1] Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
{jiaxing.huang, shijian.lu, dayan.guan}@ntu.edu.sg
[2] University of Electronic Science and Technology of China, China
zhangxiaobing@std.uestc.edu.cn

Abstract. Recent advances in unsupervised domain adaptation for semantic segmentation have shown great potentials to relieve the demand of expensive per-pixel annotations. However, most existing works address the domain discrepancy by aligning the data distributions of two domains at a global image level whereas the local consistencies are largely neglected. This paper presents an innovative local contextual-relation consistent domain adaptation (CrCDA) technique that aims to achieve local-level consistencies during the global-level alignment. The idea is to take a closer look at region-wise feature representations and align them for local-level consistencies. Specifically, CrCDA learns and enforces the prototypical local contextual-relations explicitly in the feature space of a labelled source domain while transferring them to an unlabelled target domain via backpropagation-based adversarial learning. An adaptive entropy max-min adversarial learning scheme is designed to optimally align these hundreds of local contextual-relations across domain without requiring discriminator or extra computation overhead. The proposed CrCDA has been evaluated extensively over two challenging domain adaptive segmentation tasks (*e.g.*, GTA5 → Cityscapes and SYNTHIA → Cityscapes), and experiments demonstrate its superior segmentation performance as compared with state-of-the-art methods.

Keywords: Semantic segmentation, Unsupervised domain adaptation, Contextual-relation consistent

1 Introduction

Semantic segmentation has been a longstanding challenge in computer vision, which aims to assign class labels to every pixel of an image [59]. Deep learning based approaches have achieved great successes at the price of large-scale densely-annotated datasets [3, 27, 9] which are prohibitively expensive to collect [9]. One way of circumventing this constraint is to use synthesized images with automatically generated labels (*e.g.*, synthesized [36] or game-engine produced
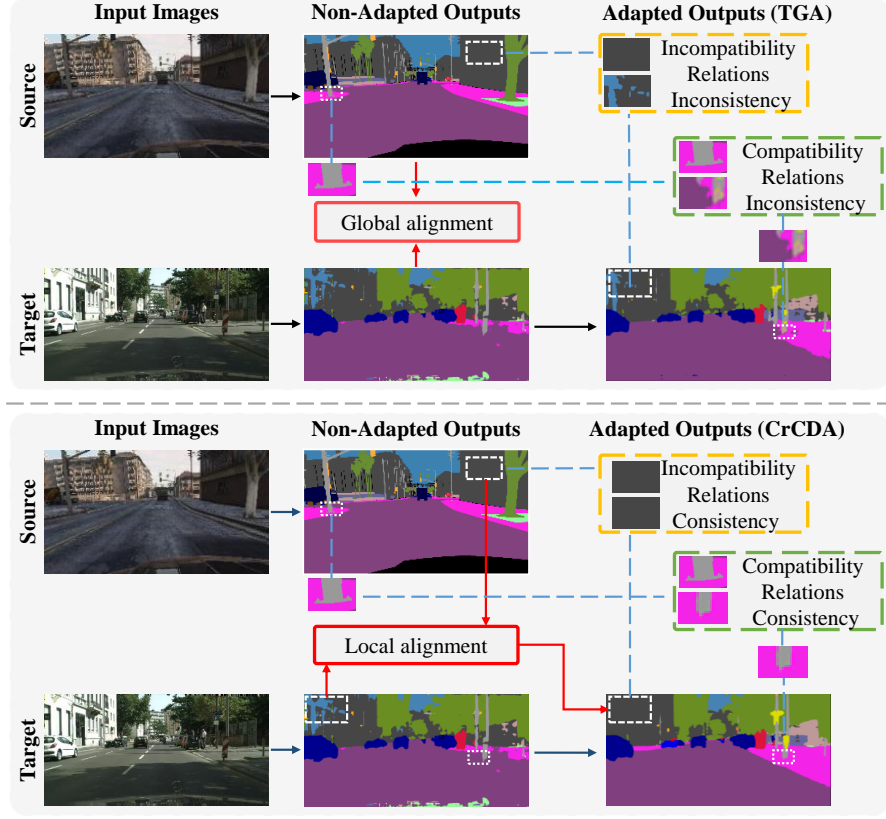
Fig. 1. Our contextual-relation consistent domain adaptation (CrCDA) improves domain adaptive semantic segmentation significantly: The traditional domain adaptive segmentation shown in the upper part employs discriminators for global alignment in the output space [44, 49, 45] (*e.g.*, probability-/entropy-/patch-represented output), which tends to introduce segmentation errors due to the neglect of local contextual consistency. Our CrCDA shown in the lower part adapts features at local level for contextual-relation consistency between the source and target domains which generates more accurate segmentation consistently. In the graph, "compatibility relations" refer to visual patterns with high co-occurrence frequency (*e.g.*, "pole" should be beside the "sidewalk"), and "incompatibility relations" refer to visual patterns with low co-occurrence frequency (*i.e.*, "sky" should not in the "building").

[35] data) in network training. Unfortunately, such models usually undergo a drastic performance drop when applied to real-world images [53] due to the domain bias and shift [39, 32, 44, 48, 40, 30].

Unsupervised domain adaptation (UDA) has been introduced to address the domain bias/shift issue. To reduce the cross-domain discrepancy, most state-of-the-art UDA methods [46, 44, 31, 49, 45, 17] exploit adversarial learning for

distribution alignment in the intermediate feature [46], output [44, 31] or latent [45, 49] space. Among this cohort of adversarial-based methods, a common and pivotal step is the employment of a discriminator [16] that predicts a domain label for data being either source or target domain. However, the discriminator works only on image-level and merely achieves global consistency (*i.e.*, locational/spatial distributions consistency), where local contextual consistency (*i.e.*, region-wise contextual-relationships) is largely neglected.

Local contextual-relationships are ubiquitous and provide important cues for scene segmentation. They can be formulated in terms of semantic compatibility/incompatibility relations between one thing/stuff and its neighbouring things/stuff. Under this formulation, a compatibility relation is an indication of visual patterns with high co-occurrence frequency, e.g. a pole beside a sidewalk, and an incompatibility relation is an indication of visual patterns with low co-occurrence frequency, e.g. a person above a driving car. The contextual information has been extensively explored in supervised semantic segmentation, whereas the local contextual-relationships is largely neglected in unsupervised domain adaptive semantic segmentation though they're beneficial in addressing local contextual consistency and inconsistency in the target domain, as illustrated in Fig. 1.

To this end, we propose an unsupervised domain adaptation method for semantic segmentation that explicitly models the local contextual-relations in the feature space of source domain (with label) and then transfers this contextual information into the target domain (without label), ultimately improving target domain segmentation quality, as shown in Fig. 1. We first establish local contextual-relationships pseudo annotations in the source domain. This can be achieved by sampling regions from pixel-level ground-truth maps of source images and clustering the sampled regions to indexed N/M groups via Dbscan [12], as illustrated in Fig. 4. With the local contextual-relationships pseudo annotations in source domain, we can train a classifier C to explicitly models/learns the local contextual-relations in the feature space of source domain, and then transfers/enforces these local contextual-relations into target domain.

Following current discriminator-based global alignment methods [44, 45, 49, 31], a intuitive idea is to employ hundreds of discriminators to align hundreds of contextual-relations across domain where a single discriminator focuses on a single contextual-relation, or employ just one discriminator to align all contextual-relations across domains. Obviously, the former is cumbersome which requires much redundant computation, while the latter is not aware of a variety of contextual-relations in the data distribution and may end up biasing to low-level/simple difference. Therefore, different from current discriminator-based global alignment methods [44, 45, 49, 31], we enforce these local contextual-relations on target domain via adaptive entropy max-minimizing (AEMM) between classifier C and feature extractor E that estimates prototypical feature representations of these local contextual-relations and congregates neighboring target incorrect samples/contextual-relations to the approximated correct source prototypes alternatively, ultimately leading to consistent local contextual-relations across do-

mains. In this way, our method requires no discriminator which is normally used in UDA-based semantic segmentation and introduces training instability and extra components. In addition, this AEMM learning scheme can also be applied into pixel-/global-scale training.

The contributions of this work can be summarized in three aspects. First, we propose an unsupervised domain adaptation method for semantic segmentation that explicitly models the local contextual-relations in the feature space of source domain (with label) and then transfers this contextual information into target domain (without label). To the best of our knowledge, this is the first effort to explore contextual information for UDA-based semantic segmentation. Second, it introduces a novel adaptive entropy max-minimizing adversarial learning scheme to effectively align hundreds of local contextual-relations across domain, which requires no discriminator and adds no overhead. Third, it shows the proposed method can be seamlessly integrated into existing domain adaptation techniques without extra overhead except two classifiers and achieves consistent improvements on semantic segmentation. Fourth, extensive evaluations over two challenging UDA tasks GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes show that our method achieves superior semantic segmentation performance consistently.

2 Related works

Current UDA-based semantic segmentation methods are threefold: adversarial learning based approach [13, 23, 24, 19, 28, 29, 46, 31, 44, 7, 11, 5, 6, 51], image translation based approach [18, 42, 50, 34, 25, 52, 54, 20, 2, 8], and pseudo-labels based approach [64, 61, 58, 15, 21].

Adversarial learning based approach: Adversarial learning based UDA has been extensively explored for semantic segmentation, where a discriminator is employed to minimize the divergences between source and target domains in feature or output spaces. [19] first applies adversarial learning for UDA based semantic segmentation by aligning feature space at global scale. Curriculum domain adaptation [55] utilizes certain inferred properties (*e.g.*, superpixel and global label distributions) as the guidance to train the segmentation network. In [44] and [7], the adversarial learning is used to align the global structure to benefit from the scene layout consistency across domains, where [7] integrates a target guided distillation module to achieve style adaptation. In addition, [38, 39] combines adversarial learning and co-training to achieve domain adaptation via maximizing the discrepancy between two classifiers' outputs.

Image translation based approach: Inspired by the recent advances in image synthesis (*e.g.*, CycleGAN [60]), a number of GAN-based methods are proposed to generate target images conditioned on the source, which can help reduce the domain discrepancy before training segmentation models. CyCADA [18] uses CycleGAN to generate target images conditioned on the source images and achieves input space adaptation with a joint adversarial learning for feature alignment. A similar method, DCAN [50], implements channel-wise feature alignment to preserve spatial structures and semantic concepts in the generator

and segmentation network. [42] transfers the information of the target domain to the learned embedding via the joint adversarial learning between generator and discriminator. Besides using GANs [16] to align the embedding across domains, [62] proposes a novel conservative loss to penalize the extremely easy and difficult cases while enhancing moderate examples.

Re-training based approach: Another approach of UDA based semantic segmentation is pseudo label re-training [64, 63, 26] that uses high-confident predictions as pseudo ground truth for the target unlabelled data to finetune the model trained on the source data. In [64], class balancing and spatial prior are included to guide the iterative re-training in target domain. [49] proposes a soft-assigned version of re-training, where it enforces the "most-confused" pixels (*e.g.*, with equal probabilities for all classes) to become more confident (*i.e.*, with either low or high probability for each class) by entropy minimization. [64] instead implements iterative learning on high-confident pixels.

Our method does not follow either global/class-wise feature space alignment using discriminators [19, 28, 29, 46, 31, 7] or re-training on target data [64, 41]. Instead, we enforce multi-scale feature space alignment via multi-scale entropy max-minimizing. To the best of our knowledge, this is the first end-to-end multi-scale UDA network that achieves competitive performance on two challenging UDA tasks.

3   Methods

In this section, we present our framework for contextual-relationships consistent domain adaptation (CrCDA): a discriminator-free adversarial training scheme between a feature extractor module and a classifier via adaptive entropy max-minimizing (AEMM) to align local contextual-relationships across domains. Fig. 2 illustrates our network architecture.

3.1   Problem Definition

We focus on the problem of unsupervised domain adaptation (UDA) in semantic segmentation. Given the source data $X_s \subset \mathbb{R}^{H \times W \times 3}$ with C-class pixel-scale segmentation labels $Y_s \subset (1, C)^{H \times W}$ (*e.g.*, stimulated images from game engines) and the target data $X_t \subset \mathbb{R}^{H \times W \times 3}$ without labels (*i.e.*, real images), our goal is to learn a semantic segmentation model G that performs well on the target dataset X_t. Current adversarial learning methods rely heavily on discriminators to align the distributions of source and target domains via two loss functions: segmentation loss on source data and adversarial loss for alignment.

However, there exists a crucial limitation for these approaches: even if perfect adaptation is achieved through a discriminator, the alignment is implemented on global level (*i.e.*, image-level), where local contextual information may be lost/deconstructed. The reason lies in that the discriminator can only implement alignment at global level, which inputs the whole map but outputs a digit to represent domain labels (*e.g.*, 0 or 1). In some cases, parts of local regions
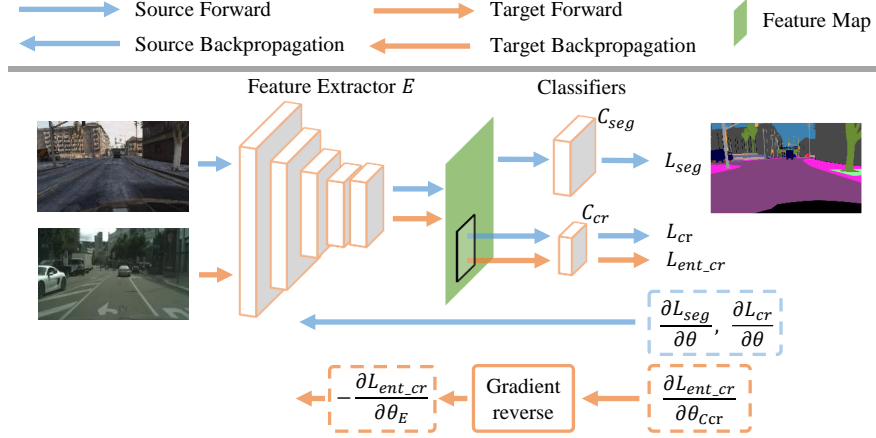
Fig. 2. Overview of our proposed contextual-relation consistent domain adaptation (CrCDA): Given images from source and target domains, the feature extractor E extracts features and feeds them to classifier C_{seg} and C_{cr} for classification at pixel and region scales. In the source flow (highlighted by arrows in blue), $\mathcal{L}_{seg}$ is computed based on the segmentation probability map from C_{seg}, $\mathcal{L}_{cr}$ is computed based on the classification probability maps from C_{cr}. In the target flow (highlighted by arrows in orange), $\mathcal{L}_{ent\_cr}$ is computed based on the classification probability maps from C_{cr}. The local-scale alignment is implemented in back-propagation by training the parts before and after the gradient reverse layer in adversarial scheme *w.r.t* $\mathcal{L}_{ent\_cr}$.

($i.e.$, local contextual-relations) have been well aligned across domains. However, the discriminator might deconstruct this existing local alignment during implementing the global marginal distribution alignment. In this paper, we define this phenomenon as "lack of local consistency ($i.e.$, local contextual inconsistency)", which is important to semantic segmentation in dense pixel-scale prediction.

3.2   Overview of Network Architecture

As shown in Fig. 2, our semantic segmentation model G consists of a feature extractor E and two classifiers ($i.e.$, C_{seg} and C_{cr}) where C_{seg} is for pixel-scale segmentation and C_{cr} is for local-scale contextual-relations learning/classification. E extracts features from input images. C_{seg} and C_{cr} classify features generated by E into pre-defined semantic classes. Specifically, C_{seg} processes features at pixel-scale, which aims to predict pixel-scale labels. The pre-defined semantic class domain for C_{seg} is the pixel-scale ground-truth, so there is no difference between C_{seg} and traditional segmentation classifier. C_{cr} processes features at local scales, which aims to predict region-scale/contextual-relations labels. The pre-defined semantic class domain for C_{cr} is the clustered contextual-relations ground-truth. The establishment procedure of clustered contextual-relations labels is described in Section 3.3 and shown in Fig. 3. We train E and the classifiers
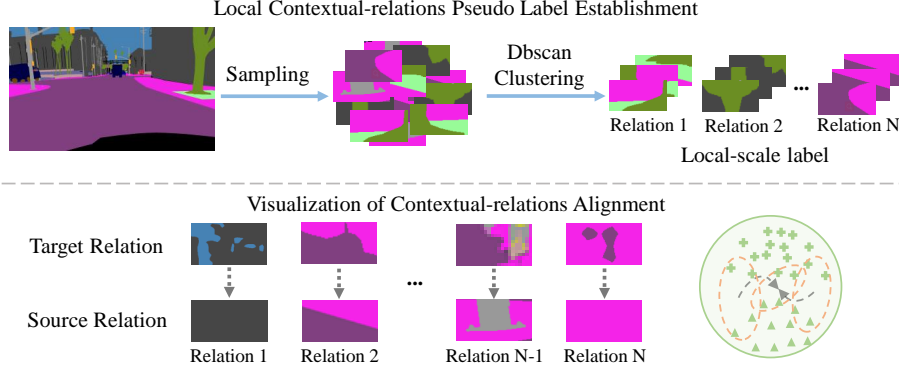
Local Contextual-relations Pseudo Label Establishment

Visualization of Contextual-relations Alignment

Fig. 3. Overview of local contextual-relation pseudo label establishment: "Dbscan clustering" means implementing Dbscan clustering based on the histogram of gradient. The effect of local contextual-relations alignment is shown at right-bottom part, with more visualization details provided in Fig. 4.

(*i.e.*, C_{seg} and C_{cr}) in an adversarial scheme to reduce domain shifts at local scales to achieve local contextual-relation consistency.

3.3   Contextual-relation consistent domain adaptation

This subsection introduces our contextual-relation consistent domain adaptation at local scales, denoted as CrCDA*, via adaptive entropy max-minimizing, as shown in Fig. 2.

   Contextual-relation pseudo label establishment. In order to implement local-scale task, we sample regions on the feature space and implement domain alignment at local scales to achieve local contextual-relation consistent domain adaptation, as shown in Fig. 3. Different from [22] that implements mode-agnostic patches alignment or [45] that aligns patch-indexed representation of the whole image only at global scales by a discriminator (*i.e.*, the probability distributions of patch index prediction of the whole images.), we aim to aligns inter-class relations within each single patch, *i.e.*, the probability distributions of pixel class prediction within each patch, *w.r.t* its mode via a classifier. Thus, the preliminary is to establish the region-scale label, where we first crop the pixel-scale ground-truth to many larger regions and then use Dbscan [12] to cluster them to assign each region a certain index label (*i.e.*, contextual-relation pseudo label). Specifically, we assign the index label to regions according to the clustering results based on the histogram of gradient. For region-scale label (*i.e.*, contextual-relation pseudo label), we cluster regions into different groups based on the histogram of gradient and assign the index label. These region-scale/contextual-relation pseudo labels can assist our network to implement alignment at local scales. Detailed information about the region-scale/contextual-relation pseudo labels is in the supplementary materials.
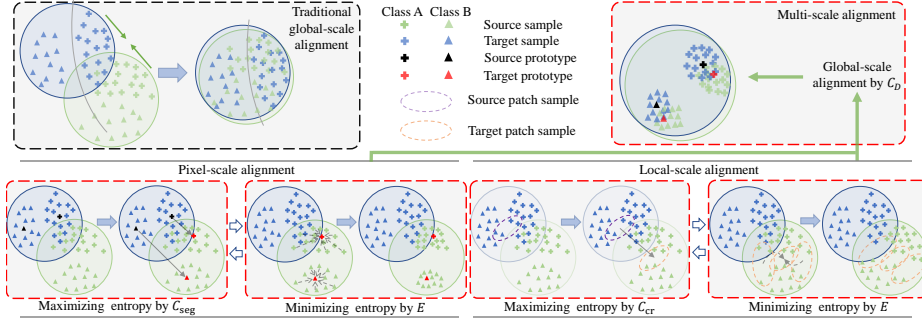
Fig. 4. Overview and comparison of the proposed AEMM at different scales: The mechanism of traditional global-scale domain adaptation is shown in the black box, where some samples are adapted into the wrong area due to the lack of local consistency (*i.e.*, local contextual-relation consistency). Our method is shown in the red boxes illustrating the alignment in pixel-scale, local-scale and global-scale. In pixel-scale alignment, C_{seg} firstly approximates the target prototypical features by maximizing entropy on target data and then E aims to congregate the features to the approximated prototypical features by minimizing entropy. Local-scale alignment works in the same scheme of pixel-scale adaptation while the only difference is the processing unit size (the former adapts a larger group of features; the latter adapts single pixel-scale features). As shown above, the global alignment is implemented by a domain classifier. Finally, the proposed AEMM can achieve feature alignment in different scales simultaneously.

Adaptive entropy max-minimizing adversarial learning scheme. In local-scale adaptation, C_{cr} aims to approximate the prototypical feature representations for each contextual-relation (*e.g.*, road-sidewalk, sky-building, pole-sidewalk, etc.) by implementing entropy maximization in target domain according to the source prototypical feature representations found via supervised learning in source domain. E focuses on extracting discriminative feature representations (near the approximated prototypical feature representations) by implementing entropy minimization. Specifically, the prototypical feature representations of source domain found with supervision are first utilized to estimate the prototypical feature representations for target data by entropy maximizing *w.r.t* C_{cr}. E then adapts the extracted feature representations to the corresponding prototypical feature representations by minimizing the entropy. The overall unsupervised domain adaptation at local scales is achieved by the adversarial training between C_{cr} and E as illustrated in Fig. 4. Different from that applied in semi-supervised learning [37], our unsupervised domain adaptation training method, referred as adaptive entropy max-minimizing (AEMM) implements entropy max-min with a regularizer $\mathcal{R}(P) = ave\{P \log P\} \times \lambda_R$ (λ_R decreases with training iteration, details are shown in appendix) for better estimating the prototypes in the target domain where no labels are available.

Source Flow. In our local-scale adaptation setting, the source data contributes to L_{seg} and L_{cr}. Given a source image $x_s \subset X_s$, its corresponding

segmentation label $y_s \subset Y_s$ and contextual-relation pseudo-label $y_{s\_cr} \subset Y_{s\_cr}$, $P_s^{(h,w,c)} = C_{seg}(E(x_s))$ is the predicted probability map $w.r.t$ each pixel over C classes; $P_{s\_cr}^{(i,j,n)} = C_{cr}(E(x_s))$ is the predicted probability map $w.r.t$ each region over N pre-defined contextual-relations classes. Therefore, it is a simple supervised learning objective to minimize L_{seg} and L_{cr}, which are expressed as:

$$\mathcal{L}_{seg}(E, C_{seg}) = \sum_{h,w} \sum_{c} -y_s^{(h,w,c)} \log P_s^{(h,w,c)} \tag{1}$$

$$\mathcal{L}_{cr}(E, C_{cr}) = \sum_{i,j} \sum_{n} -y_{s\_cr}^{(i,j,n)} \log P_{s\_cr}^{(i,j,n)} \tag{2}$$

Target Flow. As the target label is not accessible, we introduce the adversarial training scheme between feature extractor E and classifier C_{cr} to extract discriminative features for target data via adaptively max-minimizing entropy in target domain. Given a target image $x_t \subset X_t$, $P_{t\_cr}^{(i,j,n)} = C_{cr}(E(x_t))$ is the predicted probability map $w.r.t$ each region over N pre-defined contextual-relations classes. The entropy loss $L_{ent_{cr}}$ is expressed as:

$$\mathcal{L}_{ent\_cr}(E, C_{cr}) = -\frac{1}{C} \sum_{i,j} \sum_{n} max\{P_{t\_cr}^{(i,j,n)} \log P_{t\_cr}^{(i,j,n)} - \mathcal{R}(P_{t\_cr}^{(i,j,n)}), 0\} \tag{3}$$

For local-scale adaptation, we use the same back-propagation optimizing scheme with the gradient reverse layer mentioned in [57]. The training objective can be express as:

$$
\begin{aligned}
&\min_{\theta_E} \mathcal{L}_{seg} + \lambda_{cr}\mathcal{L}_{cr} + \lambda_{ent}\mathcal{L}_{ent\_cr} \\
&\min_{\theta_{C_{seg}}} \mathcal{L}_{seg} \\
&\min_{\theta_{C_{cr}}} \mathcal{L}_{cr} - \lambda_{ent}\mathcal{L}_{ent\_cr}
\end{aligned}
\tag{4}
$$

where λ_{ent} is a weight factor to control the balance of unsupervised adaptation on target data and supervised learning on source data.

3.4   CrCDA with pixel-/global-scale

This subsection introduces our CrCDA with pixel-/global-scale, denoted as CrCDA, via adaptive entropy max-minimizing, as shown in Fig. 2. Our discriminator-free AEMM adversarial training scheme can also be extended into pixel-scale and global/image-scale to form multi-scale domain adaptation.

In multi-scale adaptation, for $\mathcal{L}_{seg}$, $\mathcal{L}_{cr}$ and $\mathcal{L}_{ent\_cr}$, the objectives are the same as that in local-scale adaptation. We extend the AEMM adversarial training scheme mentioned before into pixel-scale and global-scale adaptation. For pixel-scale adaptation, we implement pixel-scale entropy loss $\mathcal{L}_{ent}$ on target data

to E and C_{seg}. For global-scale adaptation, we implement global-scale entropy loss $\mathcal{L}_{ent\_D}$ on target data to E and C_D, where C_D is a domain classifier. C_D takes the layout probability map concatenated by the two probability maps generated from C_{seg} and C_{cr} as input, and predicts domain label for it (*e.g.*, 0 for source domain, 1 for target domain). The global-alignment is achieved by the adversarial training between C_D and (E, C_{seg}, C_{cr}). Finally, our multi-scale consistent domain adaptation network is able to align domain shift at global scales, local-scale and pixel-scale simultaneously.

Similar to local-scale adaptation, we formulate the pixel-scale entropy loss as:

$$\mathcal{L}_{ent\_pix}(E, C_{seg}) = -\frac{1}{C}\sum_{h,w}\sum_{c} max\{P_{t\_pix}^{(h,w,c)} \log P_{t\_pix}^{(h,w,c)} - \mathcal{R}(P_{t\_pix}^{(h,w,c)}), 0\} \quad (5)$$

For multi-scale adaptation, we also use the same back-propagation optimizing scheme with the gradient reverse layer mentioned in [13, 14]. The training objective can be express as:

$$
\begin{aligned}
&\min_{\theta_E} \mathcal{L}_{seg} + \lambda_{cr}\mathcal{L}_{C_{cr}} + \lambda_{ent}(\mathcal{L}_{ent\_pix} + \mathcal{L}_{ent\_cr}) + \lambda_D\mathcal{L}_D \\
&\min_{\theta_{C_{seg}}} \mathcal{L}_{seg} - \lambda_{ent}\mathcal{L}_{ent\_pix} + \lambda_D\mathcal{L}_D \\
&\min_{\theta_{C_{cr}}} \mathcal{L}_{C_{cr}} - \lambda_{ent}\mathcal{L}_{ent\_cr} + \lambda_D\mathcal{L}_D \\
&\max_{\theta_{C_D}} \lambda_D\mathcal{L}_D
\end{aligned}
\quad (6)
$$

where $\mathcal{L}_D$ is provided in supplementary materials; λ_{cr}, λ_{ent} and λ_D are the weight factor to balance the unsupervised adaptation on target data and the task-specific objectives on source data.

4 Experiments

4.1 Datasets

We evaluate our unsupervised domain adaptation networks for semantic segmentation on two challenging synthesized-to-real tasks: GTA5 [35] $\rightarrow$ Cityscapes [9] and SYNTHIA [36] $\rightarrow$ Cityscapes. GTA5 contains $24,966$ synthesized images with high-resolution and we use the 19 common categories between GTA5 and Cityscapes in the same setting as in [44]. SYNTHIA contains $9,400$ synthetic images with 16 common categories in Cityscapes. We use either GTA5 or SYN-THIA as source domain. We use the unlabelled training set of Cityscapes as target domain, which includes 2975 real-world images.

4.2 Implementation Details

For a fair comparison, similar to [49] [31] [44], we utilize Deeplab-V2 architecture [3] with ResNet-101 pretrained on ImageNet [10] as our single-scale semantic

Table 1. Results of domain adaptation task GTA5 → Cityscapes. "V" means the VGG16-based model and "R" means the ResNet101-based model.

Networks	Oracle	road	side.	build.	wall	fence	pole	light	sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
FCN Wild [19]	V	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 [55]	V	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
CyCADA [18]	V	83.5	**38.3**	76.4	20.6	16.5	22.2	26.2	**21.9**	80.4	28.7	65.7	49.4	4.2	74.6	16.0	26.6	2.0	8.0	0.0	34.8
AdaptSeg [44]	V	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
CBST [64]	V	66.7	26.8	73.7	14.8	9.5	**28.3**	**25.9**	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	**18.9**	**32.4**	30.9
CLAN [31]	V	**88.0**	30.6	79.2	23.4	20.5	26.1	23.0	14.8	**81.6**	**34.5**	72.0	45.8	7.9	80.5	**26.6**	**29.9**	0.0	10.7	0.0	36.6
AdvEnt [49]	V	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
PatAlign [45]	V	87.3	35.7	79.5	**32.0**	14.5	21.5	24.8	13.7	80.4	32.0	70.5	50.5	16.9	81.0	20.8	28.1	4.1	15.5	4.1	37.5
CrCDA (ours)	V	86.8	37.5	**80.4**	30.7	18.1	26.8	25.3	15.1	81.5	30.9	**72.1**	**52.8**	**19.0**	**82.1**	25.4	29.2	10.1	15.8	3.7	**39.1**
AdaptSeg [44]	R	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
CBST [64]	R	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	**42.8**	45.9
CLAN [31]	R	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 [49]	R	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
MaxSquare[4]	R	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
PatAlign [45]	R	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
CRST [63]	R	91.0	**55.4**	80.0	**33.7**	21.4	**37.3**	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	**26.9**	26.0	42.3	47.1
CrCDA (ours)	R	**92.4**	55.3	**82.3**	31.2	**29.1**	32.5	33.2	**35.6**	83.5	34.8	**84.2**	58.9	**32.2**	84.7	**40.6**	46.1	2.1	31.1	32.7	**48.6**

segmentation network ($E + C_{seg}$). To extend our model to multi-scale network, we simply copy and modify C_{seg} to create C_{cr} and C_D with different output channels (*e.g.*,N and 1) and different output sizes due to various scales (*i.e.*, region-size and global-size). We also apply our methods on VGG-16 [43] in the same way as employing ResNet-101. Following [13] [47], a gradient reverse layer is employed to reverse the entropy loss between E and (C_{seg}, C_{cr}) during pixel-/region-scale adaptation to achieve adversarial training. The domain classifier C_D works similar to a discriminator for global-scale alignment. During training, we utilize SGD [1] to optimize our networks with a momentum of 0.9 and a weight decay of $1e-4$. The initial learning rate is set as $2.5e-4$ and decayed by a polynomial policy with a power of 0.9, as illustrated in [3]. For all experiments, the hyper-parameters λ_{ent}, λ_D, λ_{cr} and N are set as $2.5e-5$, $2.5e-5$, $5e-3$ and 100, respectively.

4.3   Comparison with state-of-art

We compare the experimental results of our method and state-of-the-art algorithms in two "Synthetic-to-real" UDA tasks with two different architectures: VGG-16 and ResNet-101. For "GTA5 → Cityscapes", we present the results in Table 1 with comparisons to the state-of-the-art domain adaptation methods [49, 31, 44, 19, 55, 18, 56, 57]. Our contextual-relation consistent domain adaptation, expressed as CrCDA, achieves comparable performance to other state-of-the-art approaches on both architectures. Compared to Adapt-SegMap (output space global alignment) [44], category-level adversarial network (output space class-wise alignment) [31] and patch-represented global alignment [45] (patch-indexed latent space alignment), CrCDA consistently brings over $+2.1\%$ mIoU improvements on ResNet-101. We reckon this gain is from our end-to-end/concurrent multi-scale alignment, which indicates that local consistency

Table 2. Results of domain adaptation task SYNTHIA → Cityscapes. "V" means the VGG16-based model and "R" means the ResNet101-based model. "mIoU" and "mIoU*" are calculated over 16 and 13 classes, respectively.

SYNTHIA → Cityscapes																			
Networks	Oracle	road	side.	build.	wall	fence	pole	light	sign	vege.	sky	pers.	rider	car	bus	motor	bike	mIoU	mIoU*
FCNs Wild [19]	V	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.2	22.1
CDA [55]	V	65.2	26.1	74.9	0.1	0.5	10.7	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	29.0	34.8
AdaptSeg [44]	V	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
CBST [64]	V	69.6	28.7	69.5	**12.1**	0.1	**25.4**	**11.9**	**13.6**	**82.0**	81.9	**49.1**	14.5	66.0	6.6	3.7	**32.4**	**35.4**	36.1
CLAN [31]	V	**80.4**	**30.7**	74.7	-	-	-	1.4	8.0	77.1	79.0	46.5	8.9	**73.8**	18.2	2.2	9.9	-	39.3
AdvEnt [49]	V	67.9	29.4	71.9	6.3	0.3	19.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	31.4	36.6
PatAlign [45]	V	72.6	29.5	77.2	3.5	0.4	21.0	1.4	7.9	73.3	79.0	45.7	14.5	69.4	19.6	7.4	16.5	33.7	39.6
CrCDA (ours)	V	74.5	30.5	**78.6**	6.6	**0.7**	21.2	2.3	8.4	77.4	79.1	45.9	**16.5**	73.1	**24.1**	**9.6**	14.2	35.2	**41.1**
AdaptSeg [44]	R	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
CLAN [31]	R	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
AdvEnt [49]	R	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	**57.9**	23.8	73.3	36.4	14.2	33.0	41.2	48.0
MaxSquare[4]	R	82.9	40.7	**80.3**	**10.2**	0.8	25.8	12.8	**18.2**	**82.5**	82.2	53.1	18.0	**79.0**	31.4	10.4	**35.6**	41.4	48.2
PatAlign [45]	R	82.4	38.0	78.6	8.7	0.6	26.0	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	40.0	46.5
CrCDA (ours)	R	**86.2**	**44.9**	79.5	8.3	0.7	**27.8**	9.4	11.8	78.6	**86.5**	57.2	**26.1**	76.8	**39.9**	21.5	32.1	**42.9**	**50.0**

(*i.e.*, local contextual-relation consistency) is very important as well as global consistency and they are complementary to each other. In Table 2, we present the adaptation result for the task "SYNTHIA → Cityscapes" and consistent improvements are observed *w.r.t* state-of-the-arts. Detailed analysis is included in next subsection.

4.4   Ablation Studies and Analysis

We analyze our proposed CrCDA with several state-of-the-art baselines. In general, both single-scale form (CrCDA*) and multi-scale form (CrCDA) achieve comparable results to all the baselines in all the settings.

As shown on the first three rows in Table 3, our pixel-scale AEMM adversarial network brings +1.4% improvements in terms of mIoU over MinEnt [49]. The reason lies in that direct entropy minimization does not take the domain gap into account while our AEMM training scheme pushes the source distribution closer to target distribution during maximizing entropy on target data.

For our CrCDA with single-scale form (CrCDA*) via AEMM, it outperforms MinEnt-based contextual-relations alignment by +1.6% on ResNet-101, as shown on the second block (row4-5) in Table 3. We reckon that these improvements are contributed by our adaptive entropy max-min training scheme which considers the domain mismatch/gap while MinEnt neglects.

Our CrCDA with multi-scale form integrating three scales' adaptation (pixel-, local- and global-scale), termed as CrCDA shown on the bottom block in Table 3, achieves state-of-the-art performances 48.6% mIoU on ResNet-101. Besides, CrCDA also outperforms all current methods by over +1.5%. Compared to "Pixel+Global", CrCDA brings +2.6% improvement in mIoU, which demonstrates that local-scale alignment is essential as well as other scales (*e.g.*, pixel-scale and global-scale). In fact, the local contextual-relation consistent adaptation loss (*i.e.*, L_{ent_cr}) penalize groups of pixels predictions to achieve local-scale

Table 3. Ablation study of the proposed contextual-relation consistent domain adaptation on GTA5-to-Cityscapes using the ResNet-101 network. All settings/methods are with "L_{seg}" (bold texts represent our methods). CrCDA* represents the contextual-relation consistent domain adaptation with only single-scale (local scale).

Method	pixel-scale Ada. L_{minent}	L_{ent}^{ours}	local-scale Ada. L_{minent}	L_{ent}^{ours}	global-scale Ada. L_{adv}	L_{patadv}	L_{advent}	L_{D}^{ours}	mIoU
Without Ada.									36.6
MinEnt [61, 49]	✓								42.4
Pixel-AEMM		✓							43.8
CrCDA*-MinEnt			✓						42.1
CrCDA*-AEMM				✓					43.7
AdaptSeg [44]					✓				41.4
PatAlign [45]						✓			41.3
AdvEnt [49]							✓		43.8
Global-AEMM								✓	44.3
Pixel+CrCDA*		✓		✓					45.6
Pixel+Global		✓						✓	46.0
CrCDA*+Global				✓				✓	46.1
CrCDA		✓		✓				✓	**48.6**

Table 4. Complementary study of the proposed contextual-relation consistent domain adaptation with local-scale to current global alignment UDA methods on GTA5-to-Cityscapes using the ResNet-101 network. All methods are default with "L_{seg}".

Method	local-level Ada. L_{ent}^{ours}	global Ada. L_{adv}	L_{patadv}	L_{advent}	L_{D}^{ours}	mIoU
CrCDA* (ours)	✓					43.7
AdaptSeg [44]		✓				41.4
PatAdv [45]			✓			41.3
AdvEnt [49]				✓		43.8
GlobalAlign (ours)					✓	44.3
AdaptSeg[44]+PatAlign[45]		✓	✓			43.2
CrCDA* + AdaptSeg[44]	✓	✓				44.8
CrCDA* + PatAlign[45]	✓		✓			44.7
CrCDA* + AdvEnt[49]	✓			✓		45.2
CrCDA*+GlobalAlign (ours)	✓				✓	**46.1**

alignment, where global-scale adaptation loss operates more on image-scale (*e.g.*, scene layout) while that of pixel-scale works on the feature representation alignment of each independent pixels. The consistent results with different settings further confirm that complementary information has been learned in different scales' adaptation. The qualitative results and visualization of feature distributions are provided in Fig. 5 and 6, which further demonstrate our conjectures mentioned above. We also provide the complementary studies to demonstrate that our local contextual-relations alignment method is complementary to most existing global-scale alignment approaches, as shown in Table 4.

5 Conclusions

In this paper, we present the local contextual-relation consistent domain adaptation (CrCDA) to address the task of unsupervised domain adaptation for seman-
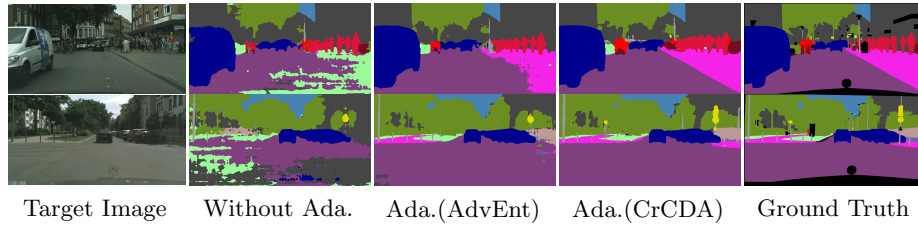
Target Image   Without Ada.   Ada.(AdvEnt)   Ada.(CrCDA)   Ground Truth

Fig. 5. Qualitative results for GTA5 → Cityscapes. Our approach (CrCDA) aligns low-level features (*e.g.*, boundaries of sidewalk, car and person *etc.*) as well as high-level features by multi-scale adversarial learning. In contrast, AdvEnt ignores low-level information because global alignment focuses more on high-level information. Thus, as shown above, CrCDA achieves both local and global consistencies while AdvEnt only achieves global consistency.

Without Adaptation   Adapted(TGA)   Adapted(CrCDA)
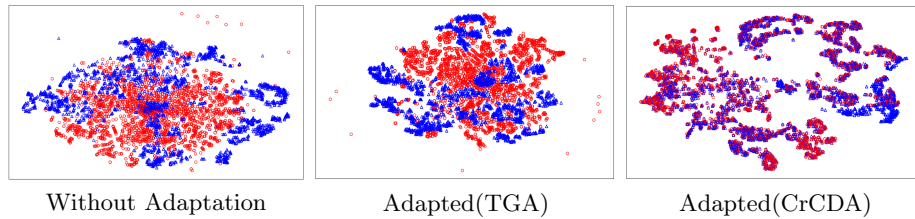
Fig. 6. Visualization of feature distributions via t-SNE [33]. "Blue": Source. "Red": Target. As shown in the first column, the feature distribution of source data is naturally more discriminative (discrete) than that of target data (uniformly distributed) due to only source supervision is available. Traditional global alignment (TGA) aligns them in global scale, where global consistency is achieved while local consistency is ignored. Thus the adapted target feature distribution is not discriminative. CrCDA aligns them with local-scale consistency (*i.e.*, local contextual-relation consistency), where both local and global consistencies are achieved. Thus the adapted target feature distribution is more discriminative and consistent with that of the source.

tic segmentation. By taking a closer look at the local inconsistency (*i.e.*, local contextual-relations inconsistency) while implementing global adaptation, Cr-CDA is able to align the domain shift in local and global scales at the same time, where local semantic consistency is normally ignored by current approaches. The experimental results on the two challenging segmentation UDA tasks validate the state-of-the-art of CrCDA.

6   Acknowledgement

References

1. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010)
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3722–3731 (2017)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
4. Chen, M., Xue, H., Cai, D.: Domain adaptation for semantic segmentation with maximum squares loss. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2090–2099 (2019)
5. Chen, Q., Liu, Y., Wang, Z., Wassell, I., Chetty, K.: Re-weighted adversarial adaptation network for unsupervised domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
6. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
7. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2018)
8. Choi, J., Kim, T., Kim, C.: Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6830–6840 (2019)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X.: Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 982–991 (2019)
12. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: Density-based spatial clustering of applications with noise. In: Int. Conf. Knowledge Discovery and Data Mining. vol. 240, p. 6 (1996)
13. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
14. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
15. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2066–2073. IEEE (2012)

16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)

17. Guan, D., Luo, X., Cao, Y., Yang, J., Cao, Y., Vosselman, G., Ying Yang, M.: Unsupervised domain adaptation for multispectral pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)

18. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)

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

20. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1335–1344 (2018)

21. Huang, J., Yuan, Z., Zhou, X.: A learning framework for target detection and human face recognition in real time. International Journal of Technology and Human Interaction (IJTHI) **15**(3), 63–76 (2019)

22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

23. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4893–4902 (2019)

24. Kang, G., Zheng, L., Yan, Y., Yang, Y.: Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–416 (2018)

25. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6936–6945 (2019)

26. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

28. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)

29. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: Advances in Neural Information Processing Systems. pp. 136–144 (2016)

30. Luo, Y., Liu, P., Guan, T., Yu, J., Yang, Y.: Significance-aware information bottleneck for domain adaptive semantic segmentation. arXiv preprint arXiv:1904.00876 (2019)

31. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: 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. pp. 2507–2516 (2019)

32. Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., Yang, Y.: Macro-micro adversarial network for human parsing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
33. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
34. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4500–4509 (2018)
35. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European conference on computer vision. pp. 102–118. Springer (2016)
36. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: 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. pp. 3234–3243 (2016)
37. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8050–8058 (2019)
38. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Adversarial dropout regularization. arXiv preprint arXiv:1711.01575 (2017)
39. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
40. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: The European Conference on Computer Vision (ECCV) (September 2018)
41. Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation. In: European Conference on Computer Vision. pp. 86–103. Springer (2018)
42. Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3752–3761 (2018)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
45. Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M.: Domain adaptation for structured output via discriminative patch representations. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1456–1465 (2019)
46. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176 (2017)
47. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
48. Vu, T.H., Choi, W., Schulter, S., Chandraker, M.: Memory warps for learning long-term online video representations. arXiv preprint arXiv:1803.10861 (2018)

49. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)

50. Wu, Z., Han, X., Lin, Y.L., Gokhan Uzunbas, M., Goldstein, T., Nam Lim, S., Davis, L.S.: Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 518–534 (2018)

51. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2272–2281 (2017)

52. Zhan, F., Huang, J., Lu, S.: Adaptive composition gan towards realistic image synthesis. arXiv preprint arXiv:1905.04693 (2019)

53. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)

54. Zhang, X., Gong, H., Dai, X., Yang, F., Liu, N., Liu, M.: Understanding pictograph with facial features: End-to-end sentence-level lip reading of chinese. In AAAI pp. 9211–9218 (2019)

55. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2020–2030 (2017)

56. Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T.: Fully convolutional adaptation networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6810–6818 (2018)

57. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)

58. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 598–607 (2019)

59. Zhu, H., Meng, F., Cai, J., Lu, S.: Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. Journal of Visual Communication and Image Representation **34**, 12–27 (2016)

60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

61. Zhu, X.J.: Semi-supervised learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2005)

62. Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D.: Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 568–583 (2018)

63. Zou, Y., Yu, Z., Liu, X., Kumar, B.V., Wang, J.: Confidence regularized self-training. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

64. Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 289–305 (2018)