Birds of a Feather Flock Together: Category-Divergence Guidance for Domain Adaptive Segmentation

Bo Yuan, Danpei Zhao, *Member, IEEE*, Shuai Shao, Zehuan Yuan, and Changhu Wang

***Abstract*— Unsupervised domain adaptation (UDA) aims to enhance the generalization capability of a certain model from a source domain to a target domain. Present UDA models focus on alleviating the domain shift by minimizing the feature discrepancy between the source domain and the target domain but usually ignore the class confusion problem. In this work, we propose an Inter-class Separation and Intra-class Aggregation (ISIA) mechanism. It encourages the cross-domain representative consistency between the same categories and differentiation among diverse categories. In this way, the features belonging to the same categories are aligned together and the confusable categories are separated. By measuring the align complexity of each category, we design an Adaptive-weighted Instance Matching (AIM) strategy to further optimize the instance-level adaptation. Based on our proposed methods, we also raise a hierarchical unsupervised domain adaptation framework for cross-domain semantic segmentation task. Through performing the image-level, feature-level, category-level and instance-level alignment, our method achieves a stronger generalization performance of the model from the source domain to the target domain. In two typical cross-domain semantic segmentation tasks, i.e., GTA5→Cityscapes and SYNTHIA→Cityscapes, our method achieves the state-of-the-art segmentation accuracy. We also build two cross-domain semantic segmentation datasets based on the publicly available data, i.e., remote sensing building segmentation and road segmentation, for domain adaptive segmentation. Our code, models and datasets are available at https://github.com/HibiscusYB/BAFFT.**

***Index Terms*— Unsupervised domain adaptation, semantic segmentation, category divergence, inter-class separation, intra-class aggregation.**

I. Introduction

SEMANTIC segmentation aims to assign a label to every pixel in the image, which normally requires large-scale pixel-level annotated data for training an applicable model. However, it is extremely time-consuming and labor-intensive to collect data with pixel-level annotations. For example, Cityscapes [2] is a widely-used benchmark dataset and it takes 1.5 hours on average to annotate an image; which sums up to about 7500 hours totally to annotate all 5000 images. However, in comparison, training an applicable semantic segmentation model on the collected data usually takes only several hours.

In recent years, photorealistic data rendered from video games and simulators with pixel-level semantic annotations have been used to train segmentation networks. Normally, the models trained on the synthetic data do not generalize well to realistic target domain. The reason lies in the different data distributions of the different domains, which is typically known as domain shift [4]. Recently, unsupervised domain adaptation (UDA) methods are proposed to address this issue. In such works, a model trained on a source domain dataset with pixel-level segmentation annotations is adapted for an unlabeled target domain. By quantifying the data distribution, domain adaptation approaches [3], [5]–[12] are proposed to minimize the feature distribution discrepancy between the source and target domains. A popular domain adaptation choice is to align the image style and feature representations of different domains [13], [14]. A majority of recent methods [15]–[17] explore semantic-level adaptation such as category-level and instance-level alignment. Among this cohort of UDA methods, a common and pivotal approach is minimizing some distance metrics between the source and target feature distributions [18]–[20]. Another effective approach, which employs GAN [21] architectures, is to minimize the accuracy of domain prediction. A GAN architecture is usually composed of a generator and a discriminator. The generator extracts features from the input images and the discriminator distinguishes which domain the features are generated from. Through a minimax game between two adversarial networks, the discriminator can thereby guide the generator to produce the target domain features with a distribution closer to that of the source domain. In recent years, the GAN-based UDA for semantic segmentation has been applied to urban scenes [13], [14], [22], aerial remote sensing images [23]–[25], LiDAR point cloud [26], [27], etc.

Although current adversarial learning methods have led to impressive results [28]–[31], there are still limitations cannot be ignored: 1) the global adversarial learning approach aligns the global feature distribution in the source and target domains

Manuscript received August 9, 2021; revised February 4, 2022; accepted March 18, 2022. Date of publication March 31, 2022; date of current version April 8, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1510905 and in part by the Air Force Equipment Pre-Research Project under Grant 303020401. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ran He. *(Corresponding author: Danpei Zhao.)*

Bo Yuan and Danpei Zhao are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: yuanbobuaa@buaa.edu.cn; zhaodanpei@buaa.edu.cn).

Shuai Shao, Zehuan Yuan, and Changhu Wang are with ByteDance AI Lab, Beijing 100086, China (e-mail: shaoshuai@acm.org; yuanzehuan@bytedance.com; wangchanghu@bytedance.com).

Digital Object Identifier 10.1109/TIP.2022.3162471

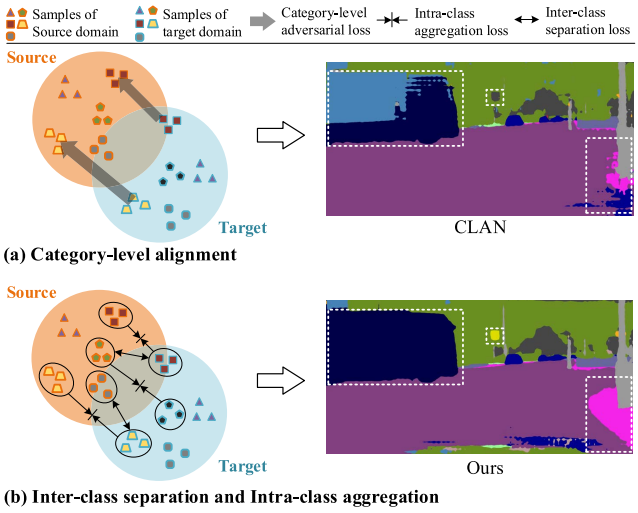


Fig. 1. Illustration of the proposed category-divergence guidance for domain adaptive segmentation. (a): Category-level alignment strategy proposed by [3]. It encourages a category-level joint distribution alignment but is confronted with the class confusion problem. (b): Our proposed inter-class separation and intra-class aggregation mechanism. Our method simultaneously performs feature alignment between the same categories and differentiating among different categories. As shown, the proposed strategy effectively reduces pixels misclassification in cross-domain segmentation task.

by training a GAN. However, when the generator can perfectly fool the discriminator, the alignment between the source and target domains is still weak for achieving a sufficient segmentation accuracy in target domain because of the low generalization on multiple categories. 2) although category-level domain adaptation approach [3] and instance-level alignment method [15] have been proposed to enhance the semantic-level alignment, there is still a problem of pixel aliasing. Specifically, the features of different categories require to be separated but the alignment strategy lacks such structural information. For example, the classes such as *sky* and *road* normally vary rarely in color, shape and position in the image, which are easily to be distinguished. While in many situations, the pixels those close to region boundaries of different categories are likely to be misclassified, as shown in Fig. 1.

To address the limitation of the traditional category-level alignment, we propose an inter-class separation and intra-class aggregation alignment strategy. By constructing a similarity measure function based on cosine distance, we conduct features alignment between the same categories and features separation among different categories across domains in the meantime. Through the measuring of the alignment complexity for each category, we design an adaptive weight to further guide the instance-level alignment. Our main contributions are summarized as follows:

- We propose a category-divergence guidance approach for cross-domain semantic segmentation. Our model efficiently reduces pixel misclassification by pulling closer feature representations of the same categories and pushing away those belonging to different categories.
- We construct a universal UDA framework from multi-level alignments including image level, feature level, category level and instance level, synergistically reducing domain gap.
- We extend the proposed UDA method to remote-sensing scenes by reforming four representative remote-sensing datasets for cross-domain building segmentation and road segmentation.
- The proposed UDA method achieves the state-of-the-art semantic segmentation accuracy on benchmark datasets including street scenes and remote-sensing images.

II. Related Works

A. Semantic Segmentation

Semantic segmentation has been significantly boosted with the development of convolutional neural networks. Since [1], the models based on fully convolutional network (FCN) [1] have grabbed massive attention. Since modeling long-range dependency information is critical for semantic segmentation, extensive efforts have been focused on increasing the receptive field through either using dilated/atrous convolutions [32], [33] or inserting attention modules [34]–[38]. Another popular path, [39]–[42] adopt encoder-decoder structures that fuse the information in low-level and high-level layers to predict segmentation mask. Reference [43] utilizes pyramid pooling to aggregate contextual information. Reference [44] starts from a high-resolution subnetwork and gradually adds high-to-low resolution subnetworks one by one to maintain high-resolution representations in the image. Ren *et al.* [45] explore neural architecture search (NAS) in semantic segmentation architecture design. Recently, [46], [47] replace traditional convolutional backbones with vision transformers. These methods consider semantic segmentation as a sequence-to-sequence prediction task to dispose the limited receptive fields. However, the advanced performance of these semantic segmentation methods often build on the large amounts of densely annotated images, which are usually difficult to collect.

B. Adversarial Learning

Generative adversarial networks (GANs) [21], [48], [49] learn two networks, i.e., a generator and a discriminator, in a staged zero-sum game fusion to generate images from inputs. The key component enabling GANs is the adversarial constraint, which makes the generated images to be indistinguishable from real images. The GAN-based methods have been widely used in image-level domain mapping. This task focuses on transferring the image style from source domain to target domain, which is popular in image-to-image translation [13], [50]–[53] and domain adaptation [6], [16], [54], [55].

C. Domain Adaptation for Semantic Segmentation

Many UDA works are designed for classification, like ADDA [56], MMD [57], *et al.* With the synthetic datasets including GTA5 [58], SYNTHIA [59], Synscapes [60] are proposed, UDA for semantic segmentation is also comes to insight. From the adaptation manner, the UDA approaches

for semantic segmentation can be divided into image-level, feature-level and label-level methods.

The image-level adaptation refers to changing the appearance of images such that images from the source domain and the target domain are more visually similar. These methods [9], [10], [13] usually transfer the color, texture, illumination and other stylization factors of images from one domain to another. Choi *et al.* [61] propose a GAN-based self-ensembling data augmentation method for domain alignment. Recently Kang *et al.* [62] propose to build the pixel-level cycle association between source and target pixel pairs and contrastively strengthen their connections to diminish the domain gap. The feature-level transferring refers to matching the extracted feature distributions between the source and target domain. Deep convolutional neural networks (CNNs) [1], [63]–[65] can extract the features from the source domain and the ones from target domain. However, due to the domain shift [3], minimizing the feature distribution discrepancy with GAN [21] structure is a common practice. Tsai *et al.* [14] propose a joint consideration of pixel and feature level adaptation. Li *et al.* [66] actively select positive source information for training to avoid negative transfer by constructing a content-consistent matching mechanism. Wu *et al.* [10] raise a channel-wise feature alignment network to close the gap of the channel-wise mean and standard deviation in CNN feature maps. Lv *et al.* [16] propose a domain-invariant interactive relation transfer strategy to align both the image-level and pixel-level information. The label-level adaptation refers to producing pseudo-labels of the target domain by utilizing the knowledge learned from the source domain, where a self-supervised learning approach [15], [17], [67]–[69] is usually used. Cai *et al.* [55] study adversarial ambivalence by revising the pseudo-labels and emerge the hard adaptation regions. Besides the single-source setting, multi-source domain adaptation [70], [71] for semantic segmentation are also studied. Tasar *et al.* [72], [73] explore domain adaptation in satellite images.

III. Preliminaries

A. Problem Setting

Given a source domain dataset with images and pixel-level annotations $\{x_i^s, y_i^s | x_i^s \in X^S, y_i^s \in Y^S\}$, and a target domain with only images $\{x_i^t | x_i^t \in X^T\}$, the goal is to train a model that can produce the pixel-level predictions $\{\hat{y}_i^t\}$ of the target domain images.

B. Segmentation and Adversarial Adaptation

We focus on training a semantic segmentation model by minimizing the discrepancy between the source and target domains. Firstly, training a model G that distills knowledge from labeled-data in order to minimize the segmentation loss in the source domain:

$$\mathcal{L}_{seg}(G) = -\sum_{i=1}^{H\times W}\sum_{k=1}^{N} y_{ik} log p_{ik} \tag{1}$$

where y_{ik} and p_{ik} represent the ground truth probability and the predicted probability of class k on pixel i, respectively. Second, an adversaries-based UDA method trains G to learn domain-invariant features by fooling a domain discriminator D which is able to distinguish samples belonging to the source or target domains. This goal is achieved by minimaxing an adversarial loss defined in Eqn (2).

$$\mathcal{L}_{adv}^{f}(G, D) = -E(log(D(G(X^S)))) - E(log(1 - D(G(X^T)))) \tag{2}$$

where $E(\cdot)$ represents statistical expectation.

IV. Method

Our model consists of a multi-level alignment framework. Specifically, we conduct the global feature-level alignment together with the proposed category-level and instance-level alignment strategies. The overall network architecture is illustrated in Fig. 2.

A. Global Feature Level Adaptation

Firstly, we use cycle-consistency [9], [74] for the unpaired image-to-image translation. This image style transferring process aims to transfer image appearance from the target domain to the source domain, which can be viewed as low-level feature alignment. To realize the global feature alignment in the output space, the images from the source and target domains are imported to a parameter-shared feature extractor. And we use the spatial layout of the source- and target-domain samples as the input of the discriminator. Following [15], we impose a traditional GAN structure on the output space [14] to globally minimize the feature distribution discrepancy between the source domain and the target domain. A discriminator D will discriminate the generated output by G. Here, the generator G is composed of a feature extractor F and a classification head C and $G = F \circ C$. We minimize the feature distribution discrepancy between the source domain and the target domain by optimizing the adversarial target function as follows:

$$\min_{G} \mathcal{L}_{adv}^{f}(G, D) = -\sum_{x_i^t \in X^T} log(1 - D(S(G(x_i^t)))) \tag{3}$$

where S is the softmax operation. While the discriminator tries to distinguish which domain the feature is formed by optimizing the discriminator target function as follows:

$$\min_{D} \mathcal{L}_{D}(G, D) = -\sum_{x_i^t \in X^T} log(D(S(G(x_i^t)))) - \sum_{x_j^s \in X^S} log(1 - D(S(G(x_j^s)))) \tag{4}$$

B. Divergence-Driven Category Level Alignment

The distribution difference of homogeneous features and the confusion of heterogeneous features constitute the key part of the domain gap. For the category-level alignment across different domains, we present an Inter-class Separation and Intra-class Aggregation (ISIA) mechanism. The key idea of the proposed ISIA is to close the feature distribution distance between the same categories and extend the feature distribution

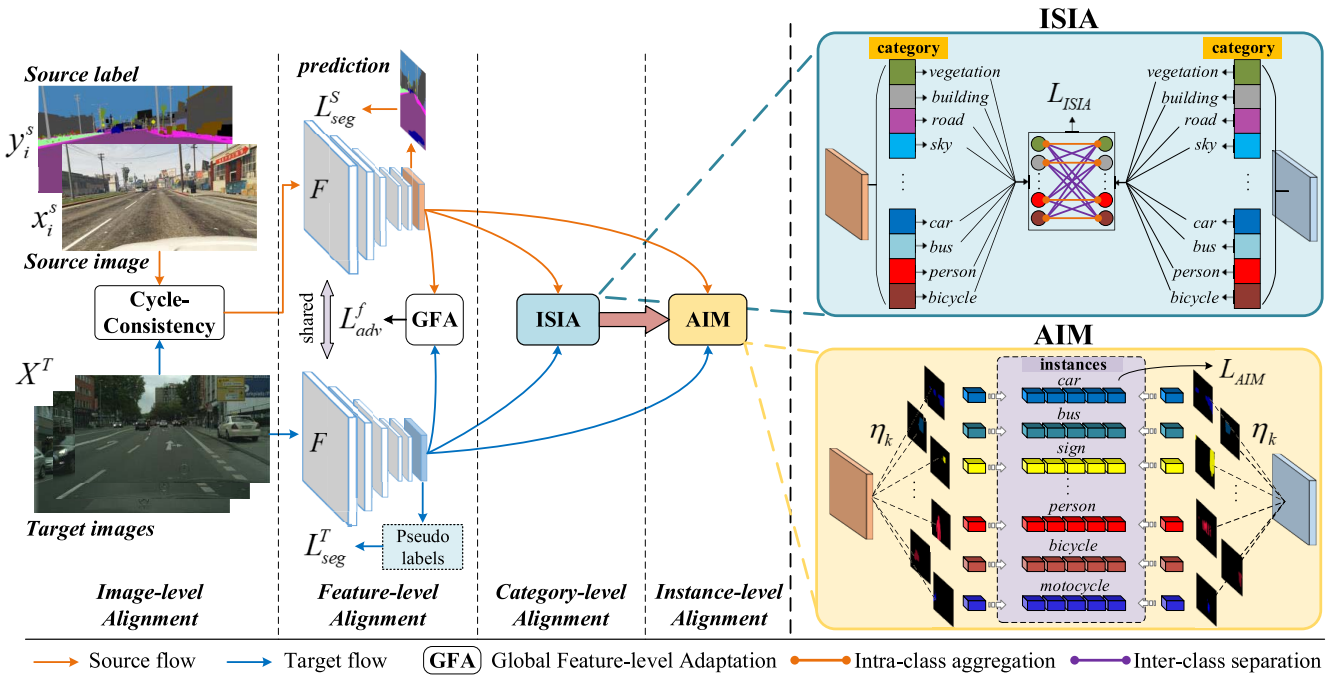


Fig. 2. Network Architecture. It consists of image-level, feature-level, category-level and instance-level alignments. GFA: global feature-level adaptation supervised by an adversarial loss. ISIA: inter-class separation and intra-class aggregation module. The features from the source and target domains are split into N classes, N is the number of semantic categories. AIM: adaptive-weighted instance matching. The foreground instances from the source and target domains are aligned by minimizing the cross-domain instance matching loss.

distance among different categories in the source and target domains.

Firstly, we feed $\hat{x}^s \in \hat{X}^S$ and $x^t \in X^T$ into a shared encoder F and two individual decoders $\{D^S, D^T\}$ to capture the features as:

$$\begin{aligned} f^s(\hat{x}^s),\, p^s(\hat{x}^s) &= D^S(F(\hat{x}^s)) \\ f^t(x^t),\, p^t(x^t) &= D^T(F(x^t)) \end{aligned} \tag{5}$$

where $\hat{x}^s$ represents the style-transferred source domain image as introduced in Sec. IV-A. $f^s(\hat{x}^s), f^t(x^t) \in \mathbb{R}^{D\times H\times W}$ are the semantic features with dimension D, $p^s(\hat{x}^s), p^t(x^t) \in \mathbb{R}^{N\times H\times W}$ are the probability predictions. In our implementation, D is set to 2048 and N_c represents the number of semantic categories. For our category-level domain adaptation, the key is to align the same category and differentiate the different categories. In high dimensional space, features are sparsely distributed. We extract $\{c_i^s, c_i^t \mid c_i \in \mathbb{R}^{1\times N_c}, i = 1, 2, \ldots, N\}$ from $\{p^s(\hat{x}^s), p^t(x^t)\}$ by selecting the corresponding channel. Thus for features those belong to the same category, our goal is to close the distance between source-domain features and target-domain features. For features those belong to different categories, the goal is to separate the feature distributions. We use cosine distance to measure the feature similarity of different categories:

$$D_{cosine}(c_i, c_j) = \frac{c_i \cdot c_j}{||c_i|| \times ||c_j||}, where \quad i \neq j \tag{6}$$

where c_i and c_j represent feature vector belonging to i-th and j-th class, respectively. Because the cosine distance ranges from -1 to 1, here we design Eqn. (7) to normalize the distance value to [0, 1] for training convenience.

$$D_{sim}(c_i, c_j) = 0.5 + 0.5 \times D_{cosine}(c_i, c_j) \tag{7}$$

Here for all categories across domains, we pull closer features those belonging to the same category and push away those belonging to different categories. Specifically, we use the L1 norm and the cosine similarity defined in Eqn. (7) to measure the embedding distance between the same and different categories, respectively. The inter-class separation and intra-class aggregation loss is defined as:

$$\mathcal{L}_{ISIA} = \sum_{i=1}^{N} ||c_i^s - c_i^t||_1 + \beta \sum_{i=1}^{N_c} \sum_{k=1,k\neq i}^{N_c} D_{sim}(c_i^s, c_k^t) \tag{8}$$

where c_i^s and c_i^t represent the feature of the i-th class of the input image belongs to the source domain and the target domain, respectively. β is used to weigh the contribution of inter-class separation during the training.

C. Category-Guided Instance Level Alignment

Reference [15] splits the objects into background stuff that usually shares similar appearance across different domains, and foreground things that often have much larger variance across images. It indicates that the foreground classes may contribute the most discrepancy across different domains. Motivated by this observation, we focus on the foreground classes those have large appearance variation and design an Adaptive-weighted Instance Matching (AIM) strategy. However, due to the lack of instance-level annotations from the source domain, we first generate the instance masks by finding

the disconnected regions for each class in the label map L follows [15]. By coarsely segmenting the intra-class semantic regions into multiple instances, the instance-level feature representations in one image is expressed as follows:

$$R_k = \{r_{k_1}, r_{k_2}, \ldots, r_{k_n}\} = \Gamma(L, k)$$
$$\mathfrak{L}(r, f) = \frac{\sum_{(h,w)} r^{(h,w)} f^{(h,w)}}{max(\epsilon, \sum_{(h,w)} r^{(h,w)})} \quad (9)$$

where r_{k_i} represents the i-th ($i \in \{1, \ldots, n\}$) binary mask of the connected region belonging to class k. Γ is the operation to find the disconnected regions of class k from the label mask L. f is the feature map generated by the feature extractor network. h and w are the height and width of the feature maps. ϵ is a regularizing term. $\mathfrak{L}$ is the operation to generate the instance-level feature representation.

Considering the category-level alignment described in Sec. IV-B, we build a ranking list to measure the complexity of the category-level adaptation across domains. We denote category-level adaptation complexity for each class as $R_{ac} = \{\zeta_k | k = 1, 2, \ldots, N_{ins}\}$, where N_{ins} is the category number of instance. ζ_k is computed by Eqn. (10).

$$\zeta_k = \frac{||c_k^s - c_k^t||_1}{max(||c_i^s - c_i^t||_1) - min(||c_i^s - c_i^t||_1)}$$
$$\eta_k = \frac{\zeta_k}{max(||\zeta_i - \zeta_j||_1)}, \quad i, j = 1, 2, \ldots, N_{ins} \quad (10)$$

where $k, i \in \{1, \ldots, N_{ins}\}$. ζ_k is updated by every batch and η_k is to avoid the weight saltus during the training. Thus the instance features across the source and target domains can be pulled closer by minimizing the cross-domain instance matching loss:

$$\mathcal{L}_{AIM} = \sum_i \sum_{k \in N_{ins}} \frac{\eta_k}{|R_k^t|} \sum_{r^t \in R_k^t} \min_j ||\mathfrak{L}(r^t, f_i^t) - s_j^k||_1 \quad (11)$$

where $i \in \{1, 2, \ldots, |X^T|\}$ and $R_k^t = \Gamma(L_{P_i}^t, k)$. s_j^k represents the j-th source domain semantic feature sample of class k. Here η_k is used to weigh the instance-level alignment of k-th class.

D. Integrated Objective

We train our model in a two-step way. Firstly, due to the lack of the target domain labels, we train our model with an initial step defined in Eqn. (12).

$$\mathcal{L}_{init} = \min_G(\lambda_{seg}\mathcal{L}_{seg}^S + \lambda_{adv}\mathcal{L}_{adv}^f + \lambda_{ISIA}\mathcal{L}_{ISIA} + \lambda_{AIM}\mathcal{L}_{AIM}) + \min_D \lambda_D \mathcal{L}_D \quad (12)$$

Then we use self-supervised learning approach same to [15] to generate pseudo labels to the pixels with high confidence of the predicted labels in the target domain training set images. Finally, we retrain our proposed models as follows:

$$\mathcal{L}_{total} = \min_G(\lambda_{seg}(\mathcal{L}_{seg}^S + \mathcal{L}_{seg}^T) + \lambda_{adv}\mathcal{L}_{adv}^f + \lambda_{ISIA}\mathcal{L}_{ISIA} + \lambda_{AIM}\mathcal{L}_{AIM}) + \min_D \lambda_D \mathcal{L}_D \quad (13)$$

where $\mathcal{L}_{seg}^S$ and $\mathcal{L}_{seg}^T$ are cross-entropy losses defined in Eqn. (1), which are used for measuring the prediction map of source domain and the target domain, respectively. λ_{seg}, λ_{adv}, λ_{ISIA}, λ_{AIM} and λ_D are the weight parameters for the losses. The pseudocode of the proposed method is shown in Algorithm 1.

Algorithm 1 Pseudocode of the Proposed Framework

Input: The source domain images and labels $\{x^s, y^s | x^s \in X^S, y^s \in Y^S\}$, the target domain images $x^t \in X^T$;
Output: Pixel-level prediction $\hat{y}_t$ of target domain images;
Suppose: segmentation network ***Seg***, *init_iters=40k*, *total_iters=120k*, adapted model $\boldsymbol{M_{step1}}$, $\boldsymbol{M_{step2}}$;
$\hat{x}^s \leftarrow CycleGAN(x^s)$, pair $\{\hat{x}^s, y^s\}$;
for *curr_iter* **in** *init_iters*:
    $\hat{y}_{pred}^s \leftarrow$***Seg***$(x^s)$, calculate $\mathcal{L}_{seg}^S$ {forward pass}
    $f^s(\hat{x}^s), p^s(\hat{x}^s) \leftarrow D^S(F(\hat{x}^s))$ {forward pass}
    $f^t(\hat{x}^t), p^t(\hat{x}^t) \leftarrow D^T(F(\hat{x}^t))$ {forward pass}
    Calculate $\mathcal{L}_{adv}^f, \mathcal{L}_D, \mathcal{L}_{ISIA}, \mathcal{L}_{AIM}$
    Optimize $\mathcal{L}_{init}$ {backward pass}
return $\boldsymbol{M_{step1}}$
Generate pseudo label $\tilde{y}^t \in Y^T$ via $\boldsymbol{M_{step1}}$
for x_i^t in X^T:
    $\tilde{y}_i^t \leftarrow \boldsymbol{M_{step1}}(x_i^t)$ {forward pass}
for *curr_iter* **in** *total_iters-init_iters*:
    $\hat{y}_{pred}^s \leftarrow$***Seg***$(x^s)$, $\hat{y}_{pred}^t \leftarrow$***Seg***$(x^t)$ {forward pass}
    Calculate $\mathcal{L}_{seg}^S, \mathcal{L}_{seg}^T$
    Repeat step 5-7
    Optimize $\mathcal{L}_{total}$ {backward pass}
return $\boldsymbol{M_{step2}}$

E. Network Architecture and Implementation

For feature extractor, we directly utilize the DeepLab-v2 [32] framework with ResNet-101 [63] pretrained on ImageNet [75] with 5 convolutional layers as the segmentation network. For discriminator network D, we adopt a similar structure with [3], which consists of 5 convolution layers with kernal 4×4 with channel numbers $\{64, 128, 256, 512, 1\}$ and stride of 2. Each convolution layer is followed by a Leaky-ReLU [76] parameterized by 0.2 negative slope between adjacent convolutional layers. The discriminator is implemented on the upsampled softmax output of the ASPP head. To train the segmentation network, we use SGD [77] as the optimizer for G with a momentum of 0.9, while using Adam [78] to optimize D with $\beta_1 = 0.9$, $\beta_2 = 0.99$. Both optimizers are set a weight decay of 5×10^{-4}. For SGD, the initial learning rate is set to 2.5×10^{-4} and decayed by a poly learning rate policy. For Adam, we initialize the learning rate to a fixed 5×10^{-5}. In the first training stage, the network is trained for $40k$ iterations by optimizing Eqn. (12). After that we further optimize Eqn. (13) for a total of $120k$ iterations. We set $\lambda_{seg} = 1$, $\lambda_D = 1$, $\lambda_{adv} = 0.001$ and batchsize as 1. All experiments are conducted on a workstation with 4 NVIDIA 2080Ti GPU cards under CUDA 11.0.

TABLE I
CROSS-DOMAIN SEMANTIC SEGMENTATION DATASETS

Type	Task	Shared classes	Spatial-resolution	Train set	Val set
Street scenes	GTA5→Cityscapes	19	-	24966	500
	SYNTHIA→Cityscapes	13	-	9400	500
Remote sensing images	MBD→IAILD	2	1.0m→0.3m	4110	800
	IAILD→MBD	2	0.3m→1.0m	2800	350
	MRD→DeepGlobe	2	1.0m→0.5m	4388	350
	DeepGlobe→MRD	2	0.5m→1.0m	500	567

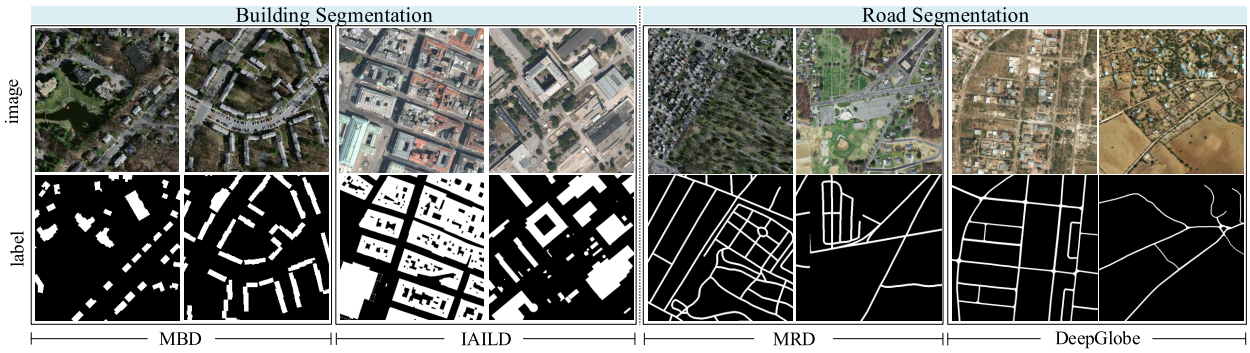


Fig. 3. The qualitative comparison of cross-domain building segmentation datasets and cross-domain road segmentation datasets.

V. EXPERIMENTS

A. Datasets

1) Street Scenes: Cityscapes [2] is a real-world dataset with 5000 street scenes of resolution 2048 × 1024. The dataset is split into training, validation and testing sets with 2975, 500, 1525 images, respectively. Following previous works [3], [15], we evaluate the models on the validation set. The Cityscapes images are resized to 1024 × 512 for both the training and testing stage. The GTA5 [58] dataset consists of 24966 fine annotated synthetic images of resolution 1914 × 1052. All the images are captured from the Grand Theft Auto V. And it shares all 19 classes with Cityscapes. SYNTHIA [59] is another synthetic image dataset that contains 9400 images of resolution 1280 × 760. Similar to [3], [15], [20], the models are evaluated on Cityscapes validation set for the 13 common classes between SYNTHIA and Cityscapes.

2) Remote Sensing Images: Domain adaptation provides a way of using the existing labeled data to run inference in unlabeled data in remote sensing image interpretation. We organize two cross-domain semantic segmentation datasets for building segmentation and road segmentation on the basis of public data, respectively. Inria Aerial Image Labeling Dataset (IAILD) [79] is a large-scale dataset for building extraction with a spatial resolution of 0.3 m and 180 labeled images with 5000 × 5000 pixels, covering different urban areas and the same areas in different time period. Massachusetts Building Dataset (MBD) [80] contains 151 sets of aerial images and corresponding single-channel label images with 2 classes. For training convenience, we randomly cut the image into 512 × 512 patches. Massachusetts Road Dataset (MRD) [80] consists of 1171 aerial images and corresponding binary label maps, each image is 1500 × 1500 pixels in size with a spatial resolution of 1 m, covering an area of 2.25 km^2. DeepGlobe [81] for road extraction contains 850 images with 2 classes annotations with size of 1024 × 1024 and the ground resolution of the image pixels is 0.5m/pixel. We also cut the images into 512 × 512 patches due to the GPU memory limitation. The cross domain datasets have difference in imaging area, object gray scale, object appearance, image annotation format, spatial resolution, etc. The datasets details are shown in Table I. Fig. 3 shows the qualitative comparison between the different domains.

We compute PASCAL VOC intersection-over-union (IoU) [83] for evaluation:

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

where TP, FP and FN are the number of true positive, false positive and false negative pixels, respectively.

B. Performance on Street Scenes

1) GTA5→Cityscapes:

a) Overall results: We compare the proposed model with the state-of-the-art UDA methods [3], [5], [11], [12], [14], [15], [17], [20], [67], [84], [86]–[89] in Table II. Our method shows strong adaptation efficiency of the model in the target domain and achieves the highest IoU in five sub-categories and the second highest IoU in another five sub-categories, especially in confusable categories like *building*, *sign* and *bus*, etc.. In terms of all categories, the proposed model achieves a new state-of-the-art performance with the mIoU of 50.7%.

TABLE II

QUANTITATIVE COMPARISON ON "GTA5→CITYSCAPES" IN TERMS OF PER-CLASS IOUS AND MIOU (%). ALL THE RESULTS ARE GENERATED FROM THE RESNET-101-BASED MODELS. THE FIRST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	road	sidewalk	building	wall	fence	pole	light	sign	vege.	terrain	sky	person	rider	car	truck	bus	train	motor	bicycle	mIoU
AdaptSeg [14]	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 [12]	89.6	58.9	78.5	33.0	22.3	41.4	48.2	39.2	83.6	24.3	65.4	49.3	20.2	83.3	39.0	48.6	12.5	20.3	35.3	47.0
CLAN [3]	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
SIBAN [84]	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
MaxSquare [85]	88.1	27.7	80.8	28.7	19.8	24.9	34.0	17.8	83.6	34.7	76.0	58.6	28.6	84.1	37.8	43.1	7.2	32.2	34.2	44.3
AdvEnt [5]	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
DPR [86]	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
PyCDA [11]	90.5	36.3	84.4	32.4	28.7	34.6	36.4	31.5	86.8	37.9	78.5	62.3	21.5	85.6	27.9	34.8	18.0	22.9	49.3	47.4
SSF-DAN [20]	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	38.8	76.9	58.8	30.7	85.7	30.6	38.1	5.9	28.3	36.9	45.4
DISE [87]	91.5	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	62.4	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
DLOW [88]	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5	42.3
FADA [17]	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
IntraDA [67]	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
Wang et al. [15]	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
ASA [89]	89.2	27.8	81.3	25.3	22.7	28.7	36.5	19.6	83.8	31.4	77.1	59.2	29.8	84.3	33.2	45.6	16.9	34.5	30.8	45.1
Ours	91.8	48.7	85.6	38.1	31.8	35.7	39.5	40.3	85.3	40.5	85.9	62.2	32.3	84.2	31.4	52.2	9.9	31.0	36.1	50.7

TABLE III

QUANTITATIVE COMPARISON ON "SYNTHIA→CITYSCAPES" IN TERMS OF PER-CLASS IOUS AND MIOU (%). ALL THE RESULTS ARE GENERATED FROM THE RESNET-101-BASED MODELS. THE MIOU COLUMN DONATED THE MEAN IOU OVER 13 CATEGORIES SHARED BY THE SYNTHIA AND CITYSCAPES. THE FIRST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	road	sidewalk	building	light	sign	vege.	sky	person	rider	car	bus	motor	bicycle	mIoU
AdaptSeg [14]	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 [3]	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 [85]	77.4	34.0	78.7	5.8	9.8	80.7	83.2	58.5	20.5	74.1	32.1	11.0	29.9	45.8
AdvEnt [5]	85.6	42.2	79.7	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	48.0
DPR [86]	82.4	38.0	78.6	3.9	11.1	75.5	84.6	53.5	21.6	71.4	32.6	19.3	31.7	46.5
FADA [17]	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
IntraDA [67]	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
Wang et al. [15]	83.0	44.0	80.3	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	52.1
ASA [89]	91.2	48.5	80.4	5.5	5.2	79.5	83.6	56.4	21.0	80.3	36.2	20.0	32.9	49.3
Ours	78.9	35.7	81.3	26.4	31.5	81.5	83.5	53.4	26.1	78.8	40.0	28.5	48.8	53.4

b) Module contributions: We first assess the contribution of each module to the overall performance in Table IV. If the model is simply trained on the source domain dataset, it achieves an mIoU of 36.6%. As introduced in Sec IV-A, we conduct image-level adaptation by transferring source image style to target domain [74] and the model achieves 42.5% mIoU. Through adversarial learning on the output space with adversarial loss proposed in [14], the mIoU is further improved to 45.3%. The IMA and GFA strategies attempt to reduce domain shift in a holistic view but ignore semantic-level information. Then we employ the proposed ISIA to train the framework and set $\lambda_{ISIA} = 0.001$ with the same weight of λ_{adv}, the model achieves an mIoU of 49.6%. Using the AIM module proposed in Sec. IV-C and setting $\lambda_{AIM} = 0.001$, the model achieves an mIoU of 50.7% by optimizing Eqn (13). Same to [15], we split the objects of *pole*, *light*, *sign*, *person*, *rider*, *car*, *truck*, *bus*, *train*, *motor* and *bike* into foreground classes and others into background classes. We focus on performing AIM on the foreground classes because they are hard to be aligned due to the large intra-class variance. Specifically, the background classes normally cover large areas and the features are easily to be distinguished. While the foreground classes usually have distinct variance in shape, texture and illuminance among instances so they are possibly to be misclassified. Fig. 6 presents the mIoU variance comparison with the increase of iteration. The proposed method shows a steadier performance and achieves a large gain compared with the global feature-level adaptation approach [14]. We further present a contrastive analysis for the feature distributions in Fig. 4. Visually, the proposed model displays higher classification accuracy in the segmentation result in such a complex scene. And from the features distribution, the proposed method can enforce intra-class features closer and the inter-class features further apart. Together with the quantitative results in Fig. 5, the proposed method can effectively improve adaptation efficiency for each category and reduce pixels misclassification especially in complex scenes.

c) Parameters study: We show the influence of β defined in Eqn. (8) to validate the contribution of inter-class separation and intra-class aggregation, respectively. As shown in Table VI, the model achieves the highest mIoU when $\beta = 1.0$. Hence we argue that the weight of β should not be either too

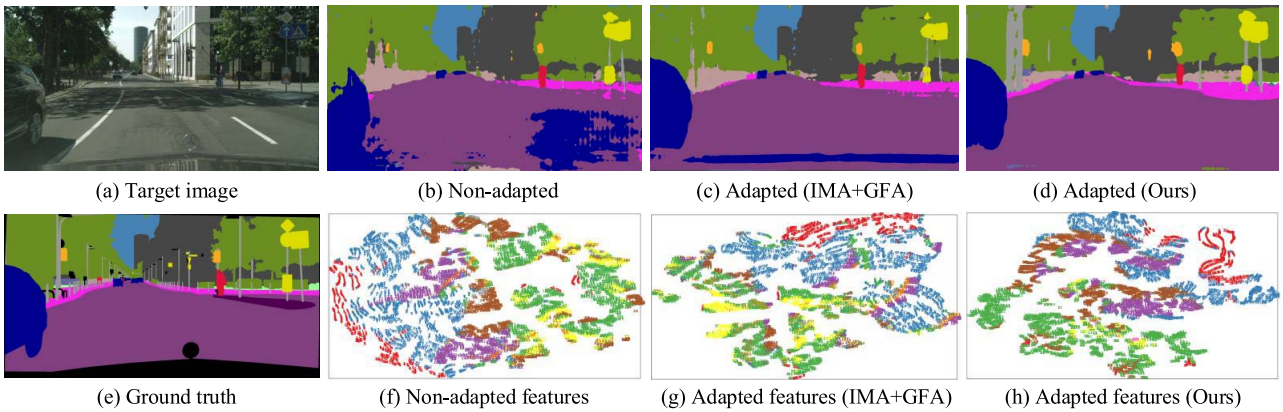


Fig. 4. Contrastive analysis of the feature distributions. (a): A target image; (b): A segmentation map of the model trained on source domain dataset only. Although the segmentation result is poor, many classes can still be correctly segmented, which indicates some classes are originally aligned without any adaptation. (c): Adapted segmentation map by adopting IMA+GFA. The segmentation performance improvement is not obvious because the IMA focuses on the appearance transferring and GFA strategy uses a simple adversarial learning in global feature output space. They lack the attention on category confusion problem. (d): Adapted result of our model. The pixels of confusable classes are well classified. Additionally, we use t-SNE [90] to map the high-dimensional features of (b), (c), (d) to 2D space shown in (f), (g), (h), respectively.

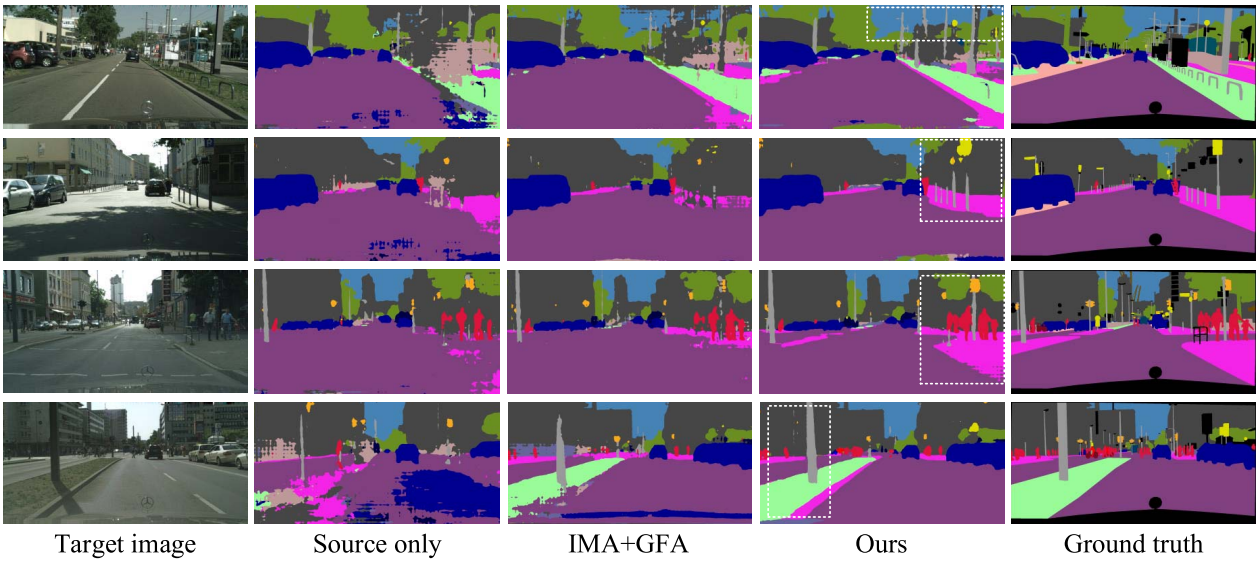


Fig. 5. Qualitative visualizations from Cityscapes validation set. For each target image, we show the corresponding non-adapted (Source only) result, the adapted result with image-level adaptation and global feature-level adaptation (IMA+GFA), the adapted result produced by our proposed model and the ground truth.

large or too small. To our best knowledge, if β is too small, the contribution of inter-class separation strategy is mild and there is high probability of pixels misclassification. While if β is too large, it leads to a drop on the segmentation accuracy. Because the influence of inter-class separation portion is violent that may override the intra-class aggregation efficiency. In our implementation, the best performance occurred when $\beta = 1.0$, which is fixed for the following experiments.

Next we discuss the contribution of the proposed ISIA by adjusting its weight coefficient λ_{ISIA} given $\lambda_{AIM} = 0.001$. As shown in Table VII, when $\lambda_{ISIA} = 0.001$, which equals to λ_{adv}, the model achieves the highest mIoU. From the experimental results, a small λ_{ISIA} may have little improvement on reducing the domain shift. A large λ_{ISIA} tends to pull the features those have large intra-class variance too much closer to the same feature sample and even aggravate the pixels misclassification, which leads to segmentation accuracy decline. By setting $\lambda_{ISIA} = 0.001$, the influence of λ_{AIM} is also explored in Table VIII. We follow [15] to adapt 10 instance features at maximum for each class from the

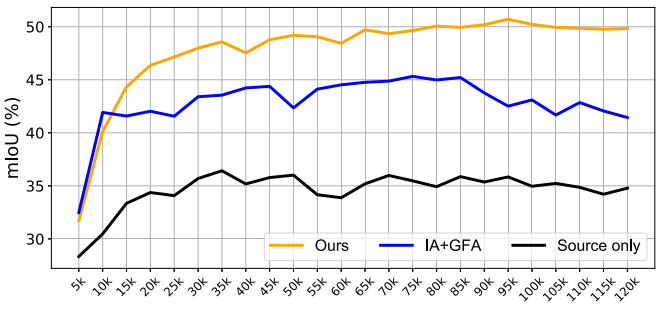


Fig. 6. mIoU comparison on Cityscapes validation set. The model is tested every $5k$ iterations on GTA5→Cityscapes.

TABLE IV

ABLATION STUDY ON GTA5→CITYSCAPES. IMA DONATES THE IMAGE-LEVEL ADAPTATION; GFA STANDS FOR GLOBAL FEATURE-LEVEL ADAPTATION; ISIA IS THE PROPOSED INTER-CLASS SEPARATION AND INTRA-CLASS AGGREGATION MECHANISM. AIM INDICATES THE PROPOSED ADAPTIVE-WEIGHTED INSTANCE MATCHING STRATEGY

Method	IMA	GFA	ISIA	AIM	mIoU(%)
Source only					36.6
+IMA [9]	✓				42.5
+GFA[14]	✓	✓			45.3
+ISIA	✓	✓	✓		49.6
+AIM	✓	✓		✓	48.8
+*all*	✓	✓	✓	✓	**50.7**
Target only					65.1

TABLE V

ABLATION STUDY ON SYNTHIA→CITYSCAPES. IMA DONATES THE IMAGE-LEVEL ADAPTATION; GFA STANDS FOR GLOBAL FEATURE-LEVEL ADAPTATION; ISIA IS THE PROPOSED INTER-CLASS SEPARATION AND INTRA-CLASS AGGREGATION MECHANISM. AIM INDICATES THE PROPOSED ADAPTIVE-WEIGHTED INSTANCE MATCHING STRATEGY

Method	IMA	GFA	ISIA	AIM	mIoU(%)
Source only					38.6
+IMA [9]	✓				42.4
+GFA [14]	✓	✓			45.6
+ISIA	✓	✓	✓		52.5
+AIM	✓	✓		✓	51.4
+*all*	✓	✓	✓	✓	**53.4**
Target only					71.7

TABLE VI

INFLUENCE OF β DEFINED IN EQN (8) ON GTA5→CITYSCAPES

β	0.1	0.5	1.0	2.0	5.0
mIoU(%)	49.8	50.3	**50.7**	50.4	50.1

TABLE VII

SENSITIVITY ANALYSIS OF λ_{ISIA} GIVEN $\lambda_{AIM} = 0.001$

GTA5→Cityscapes						
λ_{ISIA}	0.0001	0.0005	0.001	0.005	0.01	0.02
mIoU(%)	49.2	49.6	**50.7**	50.2	50.0	49.3

target domain to the source domain. Our model achieves the best performance when $\lambda_{AIM} = 0.001$. If λ_{AIM} is too small, the proposed instance-level alignment can bring a limited improvement to the model. On the other hand, if λ_{AIM} is too large, it could worsen the adaptation performance. This is because the instance features of small regions may be mixed with noisy regions due to the bottleneck of the segmentation model.

2) SYNTHIA→Cityscapes: Following the same hyper parameters discussed in Sec. V-B.1, we evaluate the proposed model on the SYNTHIA→Cityscapes task compared with [3], [5], [14], [15], [17], [67], [85], [86], [89]. As shown in Table III, our model achieves the highest mIoU with 53.4% in terms of the performance on 13 common classes.

TABLE VIII

SENSITIVITY ANALYSIS OF λ_{AIM} GIVEN $\lambda_{ISIA} = 0.001$

GTA5→Cityscapes						
λ_{AIM}	0.0001	0.0005	0.001	0.005	0.01	0.02
mIoU(%)	50.2	50.4	**50.7**	50.4	50.1	50.0

TABLE IX

ABLATION STUDY ON CROSS-DOMAIN BUILDING SEGMENTATION TASK. IMA DONATES THE IMAGE-LEVEL ADAPTATION; GFA STANDS FOR GLOBAL FEATURE-LEVEL ADAPTATION; ISIA IS THE PROPOSED INTER-CLASS SEPARATION AND INTRA-CLASS AGGREGATION MECHANISM. AIM INDICATES THE PROPOSED ADAPTIVE-WEIGHTED INSTANCE MATCHING STRATEGY

Task	Method	IMA	GFA	ISIA	AIM	Build.	Bg.	mIoU(%)
MBD ↓ IAILD	*Source only*					67.4	84.8	76.1
	+IMA [9]	✓				71.6	88.6	80.1
	+GFA [14]	✓	✓			72.9	88.9	80.9
	+ISIA	✓	✓	✓		73.3	89.2	81.3
	+AIM	✓	✓		✓	73.1	89.1	81.1
	+*all*	✓	✓	✓	✓	**73.9**	**89.6**	**81.7**
	Target only					75.1	90.2	82.6
IAILD ↓ MBD	*Source only*					35.8	87.9	61.8
	+IMA [9]	✓				39.1	87.8	63.5
	+GFA [14]	✓	✓			45.0	88.7	66.9
	+ISIA	✓	✓	✓		53.2	89.7	71.4
	+AIM	✓	✓		✓	52.7	89.3	71.0
	+*all*	✓	✓	✓	✓	**53.8**	**90.0**	**71.9**
	Target only					63.9	92.4	78.2

The contribution of each module is also analyzed on this adaptation task. From Table V, the model achieves an mIoU of 38.6% when it is trained on the source domain dataset only. The image-level adaptation brings 3.8% mIoU improvement to 42.4%. By using the GFA module, the mIoU is thereby improved to 45.6%. Here, the IA and GFA reach their performance bottleneck due to the large domain shift between the source domain and the target domain. By adding the ISIA module, the model achieves a large gain of segmentation accuracy to an mIoU of 52.5%. Using the proposed AIM can further improve the mIoU to 53.4%. Experimental results prove the proposed method has a great impact on reducing the domain shift through the efficient category-level and instance-level alignments.

C. Performance on Remote Sensing Images

To extend the proposed model to more application fields, we carry our method on cross-domain remote sensing images on two tasks, i.e., cross-domain building segmentation and cross-domain road segmentation.

1) Cross-Domain Building Segmentation: For cross-domain building segmentation, we perform the bidirectional experiments on the proposed cross-domain building segmentation

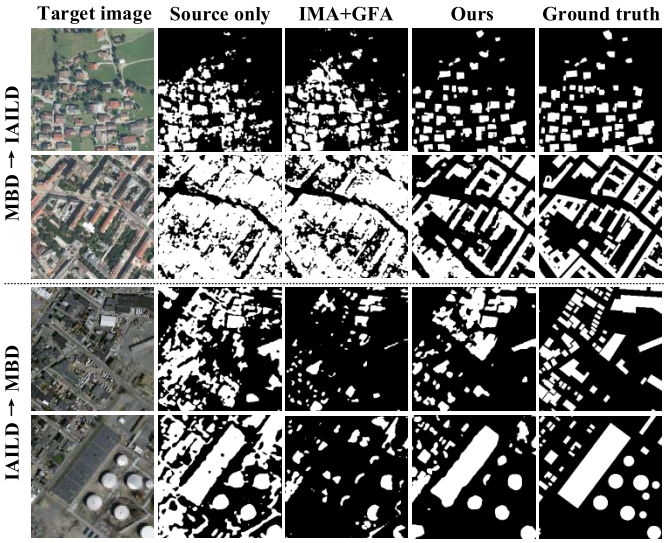


Fig. 7. Qualitative visualizations on cross-domain building segmentation task. For each target image, we show the corresponding non-adapted (Source only) result, the adapted result with image-level adaptation and global feature-level adaptation (IMA+GFA), the adapted results produced by our proposed model and the ground truth.

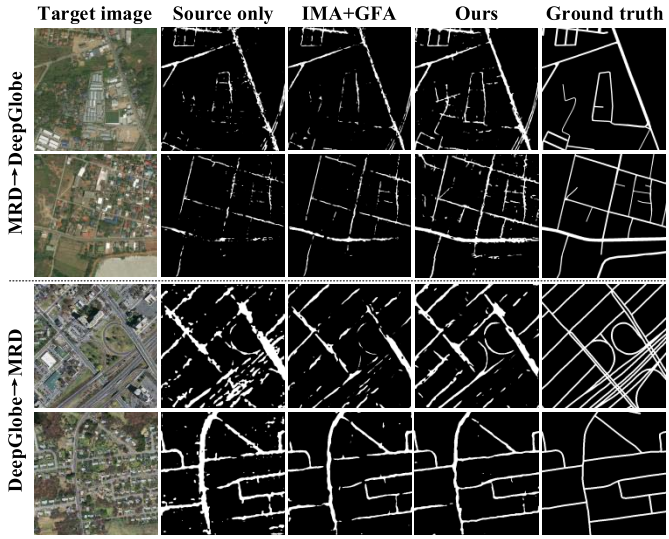


Fig. 8. Qualitative visualizations on cross-domain road segmentation task. For each target image, we show the corresponding non-adapted (Source only) result, the adapted result with image-level adaptation and global feature-level adaptation (IMA+GFA), the adapted results produced by our proposed model and the ground truth.

dataset to verify the model performance, i.e., MBD→IAILD and IAILD→MBD. In Table IX, taking MBD→IAILD as an example, compared with *Source only*'s 76.1 mIoU, our best model achieves 5.6% improvement to 81.7% mIoU. Here, it is worthy to mention that using ISIA and AIM separately can bring limited improvement compared with the model using IMA+GFA. We think it is because most of the building instances are densely arranged and vary hugely in appearance, which makes the instance extraction hard. However, by using ISIA and AIM module simultaneously, the model achieves greater performance improvement. While for the IAILD→MBD, our best model achieve 71.9% mIoU, which is 10.1% mIoU higher than the *Source only* setting. As shown in Fig. 7, we run several DA models on the target domain and output visualization results. The *Source only* model is confused on the target domain due to the large domain gap between source and target domains. Although IMA+GFA reduces domain gap from image-level and feature-level, the model's performance is still poor on account of the large amount pixel misclassification. As a comparison, the proposed model achieves a better domain adaptation effect on the target domain and effectively reduces pixel misclassification.

2) Cross-Domain Road Segmentation: For cross-domain road segmentation, we also perform the bidirectional experiments on the proposed cross-domain road segmentation dataset to verify the model performance, i.e., MRD→DeepGlobe and DeepGlobe→MRD. In Table X, taking MRD→DeepGlobe as an example, the gap between the *Source only* model the *Target only* model is 7.9% mIoU. By using the IA and GFA strategies, the adapted model achieves 3.0% mIoU improvement to 63.4% on the target domain. By using the proposed ISIA, the adapted model achieves 65.2% mIoU. However, when applying the AIM strategy on this task, the model's performance on the target domain dropped evidently. We think the reason lays on that it is hard to extract a instance for road targets, which is because they are usually connected to each other. On the other hand, since the large slenderness ratio of road targets, the down-sampling operation in the feature extraction network will cause the loss of target semantic features, which will result in poor segmentation performance. This phenomenon can be also seen in DeepGlobe→MRD task.

TABLE X

ABLATION STUDY ON CROSS-DOMAIN ROAD SEGMENTATION TASK. IMA DONATES THE IMAGE-LEVEL ADAPTATION; GFA STANDS FOR GLOBAL FEATURE-LEVEL ADAPTATION; ISIA IS THE PROPOSED INTER-CLASS SEPARATION AND INTRA-CLASS AGGREGATION MECHANISM. AIM INDICATES THE PROPOSED ADAPTIVE-WEIGHTED INSTANCE MATCHING STRATEGY

Task	Method	IMA	GFA	ISIA	AIM	Build.	Bg.	mIoU(%)
MRD ↓ DeepGlobe	*Source only*					24.8	96.1	60.4
	+IMA [9]	✓				28.5	95.8	62.1
	+GFA [14]	✓	✓			30.7	96.1	63.4
	+ISIA	✓	✓	✓		**34.2**	**96.1**	**65.2**
	+AIM	✓	✓		✓	30.4	95.8	63.1
	+all	✓	✓	✓	✓	31.8	95.9	63.9
	Target only					38.9	97.8	68.3
DeepGlobe ↓ MRD	*Source only*					30.6	93.8	62.2
	+IMA [9]	✓				33.0	95.2	64.1
	+GFA [14]	✓	✓			34.0	94.8	64.4
	+ISIA	✓	✓	✓		**37.1**	**95.2**	**66.1**
	+AIM	✓	✓		✓	36.2	94.6	65.4
	+all	✓	✓	✓	✓	36.7	95.1	65.9
	Target only					42.9	97.5	70.2

3) Comparison: We conduct comparative experiments with SOTA models [3], [14], [15] on cross-domain remote sensing datasets. As seen in Table XI and Table XII, our method achieves the highest mIoU on MBD→IAILD and MRD→DeepGlobe.

D. Ablation Study

1) Discussion on Impact of Segmentation Model: While most current DA methods for domain adaptive segmentation

TABLE XI

SEGMENTATION ACCURACY COMPARISON ON CROSS-DOMAIN BUILDING SEGMENTATION

Task	Method	mIoU(%)
MBD ↓ IAILD	Adaptseg [14]	78.5
	CLAN [3]	79.1
	Wang et al. [15]	81.1
	Ours	**81.7**

TABLE XII

SEGMENTATION ACCURACY COMPARISON ON CROSS-DOMAIN ROAD SEGMENTATION

Task	Method	mIoU(%)
MRD ↓ DeepGlobe	Adaptseg [14]	61.9
	CLAN [3]	62.6
	Wang et al. [15]	63.1
	Ours	**66.1**

use DeepLab-v2 [32] as the segmentation model. However, how the capability of segmentation model affects the domain adaptation has not been explored. By using various universally effective segmentation models, we aim to reveal the relationship between domain adaptation strategy and segmentation model performance. As seen in Table. XIII, three widely-used semantic segmentation models [32], [40], [44] are used for the domain adaptive segmentation task. For revealing the effectiveness of domain adaptation strategies, we propose a new metric called *Normalized Adaptability Measure (NAM)* as follows:

$$NAM = \frac{IoU_{Ada} - IoU_{SO}}{IoU_{TO} - IoU_{SO}} \times 100\% \tag{15}$$

where *NAM* indicates the improvement of the adapted model performance against the source only setting. Intuitively, a large *NAM* metric manifests a better adapation efficiency. *TO*, *SO* and *Ada* represent target only setting, source only setting and the adapted model, respectively.

As shown in Table XIII, we conduct three cross-domain segmentation tasks including pixel-level annotation on street scenes, remote sensing building segmentation and road segmentation, respectively. By analyzing the *NAM* metric of each segmentation model, we found that as the performance of the segmentation model improves, the improvement brought by the domain adaptation strategy will gradually increase. This shows that when the learning ability of a segmentation model is strong enough, it can also cover the domain variant to a certain extent without any other adaptation strategy. Here we propose an assumption that for the case where the difference between domains is small, the segmentation model with good performance is enough to cover most of the domain gap; for the case of large differences between domains, the segmentation model with better performance tends to overfit in the source domain, and underfit in the target domain. This is because the model will pay more attention to the different features belonging to the source domain but not the target domain. For example, we take the GTA5→Cityscapes as a *hard* domain adaptation task,

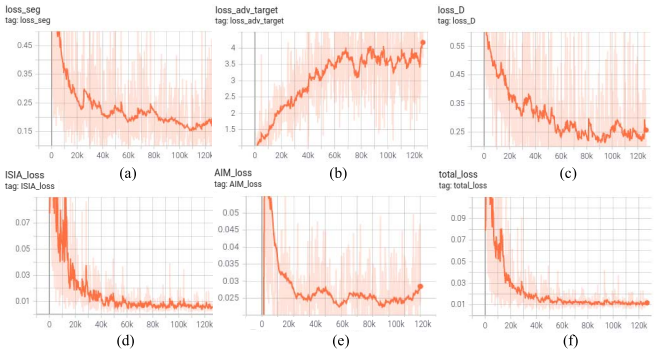


Fig. 9. Loss visualizations of the proposed model on GTA5→Cityscapes. (a) segmentation loss; (b) adversarial loss; (c) discriminant loss; (d) ISIA loss; (e) AIM loss; (f) the total loss.

TABLE XIII

IMPACT OF SEGMENTATION MODEL PERFORMANCE ON DOMAIN ADAPTATION TASK IN MEAN IOU RATE (%)

Task	Seg.-Model	Backbone	Sour. only	Ours	Tar. only	NAM
GTA5 ↓ Cityscapes	DeepLabv2 [32]	ResNet-101	36.6	50.7	65.1	48.1
	DeepLabv3+ [40]	ResNet-101	46.8	66.3	78.4	61.7
	FCN [44]	HRNet-w48	60.3	73.8	80.9	**63.7**
MBD ↓ IAILD	DeepLabv2 [32]	ResNet-101	64.2	69.1	73.9	50.5
	DeepLabv3+ [40]	ResNet-101	76.1	81.7	82.6	**85.9**
	FCN [44]	HRNetv2-w48	78.4	83.0	84.3	78.0
MRD ↓ DeepGlobe	DeepLabv2 [32]	ResNet-101	49.8	53.8	58.5	46.6
	DeepLabv3+ [40]	ResNet-101	60.4	65.2	68.3	**60.2**
	FCN [44]	HRNet-w48	65.7	71.2	75.0	59.1

because there are multiple semantic categories and large intra-class differences. As seen in Table XIII, experimental results demonstrate that when the performance of the segmentation model is enhanced, the performance improvement brought by the domain adaptation strategy is relatively strengthened since the *NAM* metric increases. While for cross-domain building segmentation task, i.e., MBD→IAILD, although the image resolutions of the source and target domains are different, the object appearance variance is small and there are only two categories, which can be taken as a *simple* domain adaptation task. Thus even the *target only* performance of FCN with HRNet-w48 [44] is better than that of DeepLab-v3+ [40], the adapted performance of *NAM* metric in the target domain is inferior than that of DeepLab-v3+. This phenomenon is also being observed in cross-domain road segmentation task, i.e., MRD→DeepGlobe. Here, it is worthy mentioning that *NAM* metric only evaluates the relative improvement of adapted model against the non-adapted model. Because the absolute performance of the adapted model increases with the performance of the segmentation model.

2) Training Stability: Since the proposed model proceeds domain adaptation on multiple levels, i.e., image-level, feature-level, category-level and instance-level, as the loss functions consists of four components, which are segmentation loss, adversarial loss, ISIA loss and AIM loss. We explore the stability of the training process. As shown in Fig. 9, the

TABLE XIV
CONTRIBUTIONS OF IS VS. IA ON CROSS DOMAIN SEGMENTATION TASK IN MEAN IOU RATE (%)

Task	IMA+GFA	+IS	+IA	+ISIA
GTA5→Cityscapes	45.3	47.4$_{+2.1}$	47.9$_{+2.6}$	49.6$_{+4.3}$
MBD→IAILD	80.9	81.1$_{+0.2}$	81.1$_{+0.2}$	81.3$_{+0.4}$
MRD→DeepGlobe	63.4	64.3$_{+0.9}$	64.6$_{+1.2}$	65.2$_{+1.8}$

TABLE XV
COMPUTATIONAL COMPLEXITY

Task	Param.	FLOPs	Memory	FPS
GTA5→Cityscapes	42.72M	183.92G	2441.78MB	12.81

segmentation loss tends to converge with iterations increasing, which indicates that the model are adapted to both the source and target domains. While the generator loss rises and discriminator loss decreases that reveals the model's feature extraction ability increases. And we see that the ISIA loss is steadily decreasing that proves the proposed ISIA strategy towards continuous optimization. While for AIM loss, it has a warm up strategy for accurate instance extracting and then the loss decreases to a small-scale fluctuating state rapidly. We think it is because the accuracy of instance extraction is limited by the segmentation model, and the instance quantity varies in different images. And for the total loss, as it is a combination of multiple losses, it towards convergence which reveal the proposed model is able to adapted to the target domain.

3) Inter-Class Separation Vs. Intra-Class Aggregation: The proposed ISIA strategy performs inter-class separation and intra-class aggregation simultaneously. To reveal the contributions of both mechanisms, we design an ablation study as shown in Table XIV. The model with the proposed ISIA is observably better than that w/o. ISIA. Among all three domain adaptive segmentation tasks, the contribution of IA is a bit greater than IS, but both can bring significant performance improvement compared to [14]. While the IS and IA can work together to achieve a better performance. As a result, we believe that pulling feature distributions of the same class across domains closer and pushing feature distributions of different classes across domains further are both beneficial to domain adaptation task.

4) Computational Complexity: Here shows the computational complexity of the proposed unsupervised semantic segmentation model. In essence, the proposed UDA method only change the distribution of parameters but does not change the FLOPs and complexity of the semantic segmentation model, we choose the representative domain adaptive segmentation task, i.e., GTA5→Cityscapes, to calculate the FPS of our method. The experimental results are shown in Table XV.

5) Failure Analysis: As seen in Fig. 10, the proposed model may show less capability in such cases: 1) Complex inter-class similarity. For example, *bus* and *truck* have the similar visual

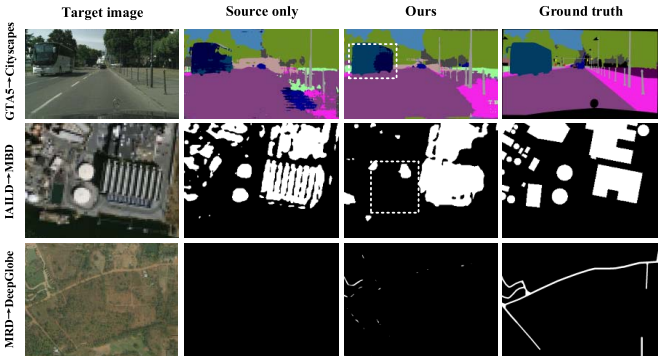


Fig. 10. Failure cases from three cross-domain semantic segmentation tasks.

appearance in GTA5→Cityscapes. Even with the proposed UDA method, there is still prediction error between these two categories limited by the feature discrimination capability of the semantic segmentation model. Of course it is worth mentioning that there is an improvement in comparison to *Source only* condition. 2) Large intra-class variation. For instance, in IAILD→MBD task, all foreground objects are labeled as one category with great difference in shape, texture, gray scale, etc. The proposed model may tend to arise pixel misclassification. 3) Bottleneck of the semantic segmentation model. We believe it may be solved by using better semantic segmentation approaches.

VI. CONCLUSION

In this paper, we propose a multi-level unsupervised domain adaptation framework for cross-domain semantic segmentation which considers category homogeneity and diversity in the meantime. Thus the model can alleviate the class confusion problem by driving intra-class features closer and inter-class features further apart. Based on the alignment complexity of each category, we design an effective instance-level alignment strategy to further enhance the adaptation validity on hard categories. Finally, the model is trained in a self-supervised way by generating the pseudo labels for the target domain. In addition, we carry out cross-domain semantic segmentation on remote sensing images to extend the domain adaptation application. This paper also explores the impact of segmentation model performance on domain adaptation efficiency. The experimental results prove the proposed method can effectively reduce pixels misclassification among confusable categories and achieve a new state-of-the-art segmentation accuracy on benchmark datasets. In the future work, the following work will be scheduled. On the one hand, the proposed UDA method can be embedded more semantic segmentation models. On the other hand, more cross domain semantic segmentation tasks are being explored. And we also attempt to extend the UDA to more complex open-world problems.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[3] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2502–2511.

[4] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statist. Planning Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[5] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[6] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. ICML*, 2018, pp. 1989–1998.

[7] G. Kang, L. Zheng, Y. Yan, and Y. Yang, "Deep adversarial attention alignment for unsupervised domain adaptation: The benefit of target expectation maximization," in *Proc. ECCV*, Sep. 2018, pp. 401–416.

[8] F. Zhu, L. Zhu, and Y. Yang, "Sim-real joint reinforcement transfer for 3D indoor navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11380–11389.

[9] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6929–6938.

[10] Z. Wu *et al.*, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. ECCV*, Sep. 2018, pp. 518–534.

[11] Q. Lian, L. Duan, F. Lv, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6757–6766.

[12] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. ECCV*, Sep. 2018, pp. 289–305.

[13] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.

[14] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[15] Z. Wang *et al.*, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12632–12641.

[16] F. Lv, T. Liang, X. Chen, and G. Lin, "Cross-domain semantic segmentation via domain-invariant interactive relation transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4333–4342.

[17] H. Wang, T. Shen, W. Zhang, L. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," 2020, *arXiv:2007.09222*.

[18] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015, *arXiv:1502.02791*.

[19] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. ECCV Workshops*, 2016, pp. 443–450.

[20] L. Du *et al.*, "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 982–991.

[21] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–9.

[22] M. Biasetton, U. Michieli, G. Agresti, and P. Zanuttigh, "Unsupervised domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1211–1220.

[23] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, p. 1369, 2019.

[24] B. Benjdira, A. Ammar, A. Koubaa, and K. Ouni, "Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks," *Appl. Sci.*, vol. 10, pp. 1–24, 2020.

[25] L. Shi, Z. Wang, B. Pan, and Z. Shi, "An end-to-end network for remote sensing imagery semantic segmentation via joint pixel- and representation-level domain adaptation," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1896–1900, Nov. 2021.

[26] P. Jiang and S. Saripalli, "LiDARNet: A boundary-aware domain adaptation model for LiDAR point cloud semantic segmentation," 2020, *arXiv:2003.01174*.

[27] L. Yi, B. Gong, and T. Funkhouser, "Complete & label: A domain adaptation approach to semantic segmentation of LiDAR point clouds," 2020, *arXiv:2007.08488*.

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

[29] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.

[30] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NIPS*, 2016, pp. 469–477.

[31] D. D. Mauro, A. Furnari, G. Patanè, S. Battiato, and G. M. Farinella, "SceneAdapt: Scene-based domain adaptation for semantic segmentation using adversarial learning," *Pattern Recognit. Lett.*, vol. 136, pp. 175–182, Aug. 2020.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[33] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[35] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[36] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.

[37] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[38] H. Zhao *et al.*, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. ECCV*, Sep. 2018, pp. 267–283.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 833–851.

[41] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5168–5177.

[42] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1857–1866.

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[44] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[45] P. Ren *et al.*, "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–34, May 2022, doi: 10.1145/3447582.

[46] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[47] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," 2021, *arXiv:2103.14030*.

[48] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. NIPS*, 2015, pp. 1486–1494.

[49] A. Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with PixelCNN decoders," in *Proc. NIPS*, 2016, pp. 4797–4805.

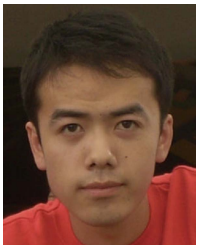
[50] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, 2016, pp. 2234–2242.

[51] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stereoscopic neural style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6654–6663.
[52] A. Royer *et al.*, "XGAN: Unsupervised image-to-image translation for many-to-many mappings," 2017, *arXiv:1711.05139*.
[53] L. Sheng, Z. Lin, J. Shao, and X. Wang, "Avatar-Net: Multi-scale zero-shot style transfer by feature decoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8242–8250.
[54] M. Gong, K. Zhang, B. Huang, C. Glymour, D. Tao, and K. Batmanghelich, "Causal generative domain adaptation networks," 2018, *arXiv:1804.04333*.
[55] B. Cai, H. Fu, R. Jia, B. Zhao, H. Li, and Y. Xu, "Exploiting diverse characteristics and adversarial ambivalence for domain adaptive segmentation," in *Proc. AAAI*, 2021, pp. 1–9.
[56] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.
[57] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. NIPS*, 2016, pp. 136–144.
[58] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," 2016, *arXiv:1608.02192*.
[59] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.
[60] M. Wrenninge and J. Unger, "Synscapes: A photorealistic synthetic dataset for street scene parsing," 2018, *arXiv:1810.08705*.
[61] J. Choi, T. Kim, and C. Kim, "Self-ensembling with GAN-based data augmentation for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6829–6839.
[62] G. Kang, Y. Wei, Y. Yang, Y. Zhuang, and A. G. Hauptmann, "Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation," in *Proc. NeurIPS*, 2020, pp. 3569–3580.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[65] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[66] G. Li, G. Kang, W. Liu, Y. Wei, and Y. Yang, "Content-consistent matching for domain adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 440–456.
[67] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3763–3772.
[68] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Proc. NeurIPS*, 2019, pp. 435–445.
[69] Y. Luo, Z. Wang, D. Huang, N. Ge, and J. Lu, "Get away from style: Category-guided domain adaptation for semantic segmentation," 2021, *arXiv:2103.15467*.
[70] S. Zhao *et al.*, "Multi-source domain adaptation for semantic segmentation," 2019, *arXiv:1910.12181*.
[71] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," 2021, *arXiv:2103.04717*.
[72] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "StandardGAN: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 747–756.
[73] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "DAugNet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1067–1081, Feb. 2021.
[74] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[75] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[76] A. L. Maas *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1. Citeseer, 2013, p. 3.
[77] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.
[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.
[79] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.
[80] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Graduate Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.
[81] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
[82] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. I–3, pp. 293–298, Jul. 2012.
[83] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.
[84] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6777–6786.
[85] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.
[86] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.
[87] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1900–1909.
[88] R. Gong, W. Li, Y. Chen, and L. Van Gool, "DLOW: Domain flow for adaptation and generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2472–2481.
[89] W. Zhou, Y. Wang, J. Chu, J. Yang, X. Bai, and Y. Xu, "Affinity space adaptation for semantic segmentation across domains," *IEEE Trans. Image Process.*, vol. 30, pp. 2549–2561, 2021.
[90] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

Bo Yuan received the B.S. and M.S. degrees from Beihang University in 2019 and 2022, respectively, where he is currently pursuing the Ph.D. degree. From 2020 to 2021, he was a Research Intern at ByteDance AI Lab. His research interests include image processing, computer vision, life-long learning, and their application in remote sensing images.

Danpei Zhao (Member, IEEE) received the Ph.D. degree in optical engineering from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006. From 2006 to 2008, she was at Beihang University for postdoctoral research. From 2014 to 2015, she was working at the Department of Computer Science, Rutgers, The State University of New Jersey, USA, as a Visiting Scholar. She is currently an Associate Professor and a Ph.D. Supervisor at the Department of Aerospace Information Engineering, Beihang University. Her research interests include saliency detection, target detection and recognition, image understanding, and their application in remote sensing images. She serves as a Standing Member for the Executive Council of Beijing Society of Image and Graphics.

Shuai Shao received the B.S. degree from the Tang Aoqing Honors Program in Computer Science, Jilin University, China. He is currently a Researcher at ByteDance AI Lab. Before joining ByteDance, he worked as a Researcher at Megvii Research from 2017 to 2020. He has published four academic papers on top-tier international conferences and two datasets.



Zehuan Yuan received the B.S. and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University, China. He is currently a Researcher at ByteDance AI Lab. He has published more than 20 academic papers on the top-tier international journals and conferences, such as ICML, CVPR, ICCV, ICLR, IJCAI, AAAI, and ECCV. His research interests lie in computer vision and machine learning. He was a Reviewer or a PC Member of CVPR, ICCV, ECCV, IJCAI, AAAI, ICML, and ICLR.

Changhu Wang received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2004 and 2009, respectively. He is currently the Head of vision technology at ByteDance, Beijing, China. Before joining ByteDance, he worked as a Lead Researcher at Microsoft Research Asia from 2009 to 2017. He worked as a Research Engineer at the Department of Electrical and Computer Engineering, National University of Singapore, in 2008.