BAPA-Net: Boundary Adaptation and Prototype Alignment for Cross-domain Semantic Segmentation

Yahao Liu Jinhong Deng Xinchen Gao Wen Li* Lixin Duan
Data Intelligence Group, University of Electronic Science and Technology of China
{lyhaolive, jhdeng1997, gxc0327, liwenbnu, lxduan}@gmail.com

Abstract

Existing cross-domain semantic segmentation methods usually focus on the overall segmentation results of whole objects but neglect the importance of object boundaries. In this work, we find that the segmentation performance can be considerably boosted if we treat object boundaries properly. For that, we propose a novel method called BAPA-Net, which is based on a convolutional neural network via Boundary Adaptation and Prototype Alignment, under the unsupervised domain adaptation setting. Specifically, we first construct additional images by pasting objects from source images to target images, and we develop a so-called boundary adaptation module to weigh each pixel based on its distance to the nearest boundary pixel of those pasted source objects. Moreover, we propose another prototype alignment module to reduce the domain mismatch by minimizing distances between the class prototypes of the source and target domains, where boundaries are removed to avoid domain confusion during prototype calculation. By integrating the boundary adaptation and prototype alignment, we are able to train a discriminative and domain-invariant model for cross-domain semantic segmentation. We conduct extensive experiments on the benchmark datasets of urban scenes (i.e., GTA5→Cityscapes and SYNTHIA→Cityscapes). And the promising results clearly show the effectiveness of our BAPA-Net method over existing state-of-the-art for cross-domain semantic segmentation. Our implementation is available at https://github.com/manmanjun/BAPA-Net.

1. Introduction

Because of the powerful representation ability of deep convolutional neural networks [26], it has deeply boosted the performance of the computer vision tasks including image recognition [45, 21], object detection [16, 32], semantic segmentation [33, 63, 3], etc. They all require plentiful images and accurate annotations to train high-performance models. Compared with image recognition, semantic segmentation is more complex and aims at classifying each pixel in an image. Therefore, collecting the annotations for segmentation is an extremely expensive and laborious process (*e.g.*, 90 minutes per image for Cityscapes [7]).

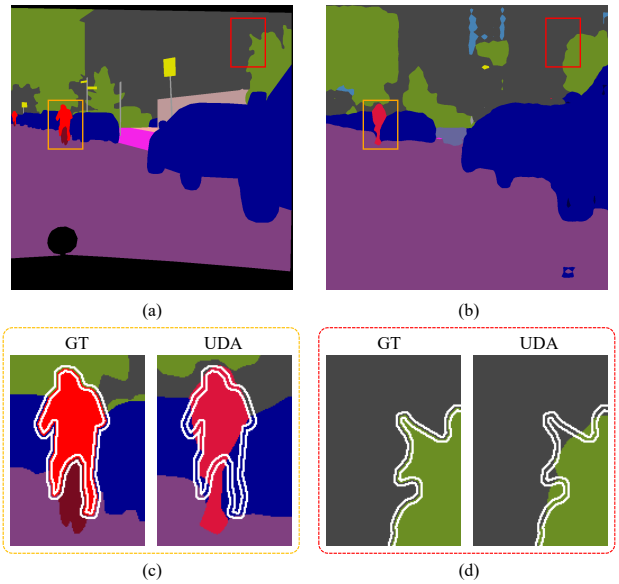


Figure 1. Comparisons between (a) the ground truth annotations and (b) the segmentation result of an existing unsupervised domain adaptation method [57] of a sample target image. It is obvious that the segmentation results of boundary pixels are worse than the inner pixels (*e.g.*, the inner part of the rider in (c) and the vegetation in (d) are predicted perfectly, while the boundary pixels are not).

A natural alternative is to collect the well-annotated synthetic data from the simulation platform where it can automatically render the various scenes (*e.g.*, sunny, rain, foggy street) with a much lower cost. For example, [43] builds a large-scale urban scene dataset obtained from the GTA5 video game. However, the trained model on such synthetic data will suffer from a significant performance drop as there exists a considerable domain discrepancy between

*The corresponding author

source and target domains. A variety of unsupervised domain adaptation methods are proposed to maximally eliminate the domain discrepancy through adversarial feature learning [22, 6, 13], entropy minimization [52, 53, 5], self-training [65, 66, 31, 11], etc.

However, we observe that current state-of-the-art methods often focus on the overall segmentation results of whole objects but neglect the importance of object boundaries. Taking the Fig. 1 as an example, some pixels along the boundaries of the person and tree are wrongly classified. The reason is that the near-boundary pixels and inner-object pixels are different, as the receptive field of the boundary sample might contain pixels from other classes, making near-boundary pixels difficult to classify. This becomes even worse in the Unsupervised Domain Adaptation (UDA) scenario, where a considerable distribution mismatch exists between source and target domains.

Therefore, the segmentation performance can be considerably boosted if we treat object boundaries properly. To achieve this, we propose a novel method called Boundary Adaptation and Prototype Alignment Network (BAPA-Net). Specifically, we first construct additional images by pasting objects from source images to target images, and we develop a so-called boundary adaptation module to weigh each pixel based on its distance to the nearest boundary pixel of those pasted source objects. Moreover, we propose another prototype alignment module to reduce the domain mismatch by minimizing distances between the class prototypes of the source and target domains, where boundaries are removed to avoid domain confusion during prototype calculation. By integrating the boundary adaptation and prototype alignment, we are able to train a discriminative and domain-invariant model for cross-domain semantic segmentation. The proposed method outperforms the state-of-the-art counterparts by a large margin on the benchmarks of GTA5→Cityscapes and SYNTHIA→Cityscapes respectively, which verifies the effectiveness of our BAPA-Net.

The main contributions of our work can be summarized as follows:

- We reveal a critical finding that existing cross-domain semantic segmentation methods neglect the importance of object boundary. Thus we propose a novel approach called Boundary Adaptation and Prototype Alignment (BAPA-Net) to maximally exploit the boundary properly.
- We develop a so-called boundary adaptation module to weigh each pixel around the boundary and a new prototype alignment module to build more reliable prototypes by removing the domain confused boundaries so that the domain mismatch between source and target domains can be reduced effectively.
- We conduct extensive experiments on the benchmark settings for urban scenes (*i.e.*, GTA5→Cityscapes and SYNTHIA→Cityscapes), and the experimental results demonstrate the effectiveness of our proposed method.

2. Related Work

Unsupervised domain adaptation. Conventional machine learning algorithms rely on a hypothesis that the training and testing data are drawn from the same distribution. However, this hypothesis typically does not hold in practice. The unsupervised domain adaptation methods [1, 23] are proposed to address this issue by eliminating the domain discrepancy between source and target domains. Many previous works try to minimize the Maximum Mean Discrepancy [20, 36, 34], KL-divergence [47], optimal transport distance [8, 9], etc. Recent methods aim to improve the domain adaptability of deep neural networks via adversarial training [35, 62]. Other methods include subspace alignment [17], geodesic flow kernel [18], transfer multiple kernel learning [14], etc.

Semantic segmentation. As a pixel-level prediction task, semantic segmentation energizes many visual applications such as medical diagnosis, autonomous driving, security. In 2014, Long [33] proposed the Fully Convolutional Network (FCN) that replaces the fully connected layer by fully convolutional layer so that the network can directly output the segmented mask map. With the development of dilated convolution [12], Deeplab [3, 4] and PSPNet [63] propose to capture more contextual information of images through multi-scale feature fusion. Recently, the research community pays more attention to the construction of the context-based attention mechanism and computational efficiency, such as RANet [56], EMANet [29]. In addition, the fine-grained segmentation of category boundaries [2, 28] has become one of the most challenging difficulties in current semantic segmentation tasks. In this work, we utilize DeepLab V2 [3] with ResNet101 [21] as our baseline architectures for semantic segmentation.

Cross-domain semantic segmentation. Due to the annotations for semantic segmentation are costly and non-trivial to acquire by human labor, how to use models trained on simulated images to achieve good performance in real scenes has gradually become a hot research topic. To address the domain shift problem, style transfer [15, 19, 42] has been used to align domain distribution on input layer (*i.e.*, pixel space). Besides, GAN [50, 38, 37, 13, 6, 54, 25] related works have been used to align domain distribution on feature and output space, respectively. [19] proposes a progressive adaptation method to alleviate the domain shift by controlling the degree of style translation from the source domain to the target domain. [50] leverages a discriminator to distinguish the outputs from the segmentation network with different domain inputs so that the

Figure 2. Illustration of the CutMix operation. The orange parts are the pasted pixels from the source image. In this work, we cut all the pixels of half classes from the labeled source images X_s, and paste them to the unlabeled target image X_t to construct the mixed image X_m.

model can minimize the domain discrepancy. [39, 40] use the image-level category information about target image to construct curriculums for the domain adaptation problem. In addition, some methods originating from semi-supervised learning are also used to address this problem, such as entropy minimization [52, 53], self-supervised training [65, 66, 31, 64, 54, 58, 24, 41, 27]. Recently, [49] utilizes the CutMix data augmentation method to address the cross-domain semantic segmentation. Different from those works, we provide a new perspective from the boundary adaptation and prototype alignment to address the cross-domain semantic segmentation problem.

3. Methodology

In this section, we explain our proposed approach BAPA-Net in detail. Given a labeled source domain and an unlabeled target domain, our goal is to learn a robust semantic segmentation model that works well not only for the source domain but also for the target domain. Formally, let us denote $\mathcal{D}_S = \{\mathcal{X}_s, \mathcal{Y}_s\}$ as the source domain training samples, where $\mathcal{X}_s = \{X_s^i|_{i=1}^{N_s}\}$ and $\mathcal{Y}_s = \{Y_s^i|_{i=1}^{N_s}\}$, each X_s is a source image, $Y_s \in \mathcal{C}, \mathcal{C} = \{1, ..., C\}$ is the corresponding pixel-level annotation with C being the number of class. Accordingly, we denote $\mathcal{D}_T = \{\mathcal{X}_t\}$ as the target domain, and $\mathcal{X}_t = \{X_t^i|_{i=1}^{N_t}\}$ where each X_t is a target image for which the annotation is unavailable.

A good cross-domain semantic segmentation model should be both discriminative and domain-invariant. Good discriminability means the model is able to distinguish samples from different classes, especially for boundary samples. And the domain-invariant ability ensures the model performs well on both domains.

For this purpose, we propose a Boundary Adaptation and Prototype Alignment Network (BAPA-Net) to enhance the segmentation model on both abilities. On the one hand, we propose a boundary adaptation approach, in which we generate additional boundary samples and enhance the model discriminability. On the other hand, we design a prototype adaptation approach to minimize the class prototype centroids of two domains in the feature space in order to learn domain-invariant features.

Simultaneously improving both abilities in one model is not easy, since they often depend on each other. For example, the generated boundary samples should also be domain-invariant. Otherwise, the segmentation model might be biased. Also, when aligning the prototypes, since the target domain images are unlabeled and only pseudo-labels can be used, the unreliable boundary samples should be excluded. We present the details of our BAPA-Net on addressing these issues in the following. The overall structure is shown in Fig. 3.

3.1. Boundary Adaptation

As aforementioned, correctly predicting the boundary samples is challenging for the cross-domain semantic segmentation task. Therefore, we propose to enhance the model's discriminability by enforcing the model to focus more on the boundary samples.

However, in the cross-domain semantic segmentation task, only images in the source domain are annotated, and the images in the target domain are totally unlabeled. While we are able to find boundary samples in source images according to their labels, it is not desirable to enforce the model to focus on these samples, since this would inevitably make the model to be biased to the source domain. To this end, we propose to employ the recent proposed CutMix method [60] to generate domain-mixed boundary samples.

Domain-mixed boundary sample generation. The CutMix employs a cut-paste data augmentation strategy for semantic segmentation. In particular, they expand the training set by randomly mixing a labeled image and an unlabeled image. At each time, they cut all the pixels of some random classes from the labeled images, and paste them to the unlabeled image, and append the mixed image into the training set to train the segmentation model.

Formally, given a labeled source image X_s and an unlabeled target image X_t, let us denote M_{mask} as the selection indicator for the pixels of randomly selected half classes in X_s, where $M_{mask}^{(h,w)} = 1$ if the pixel, located at the h-th row and w-th column, belongs to the selected classes, and $M_{mask}^{(h,w)} = 0$ otherwise. The mixed image can be presented as:

$$X_m = M_{mask} \odot X_s + (1 - M_{mask}) \odot X_t, \qquad (1)$$

where $\odot$ is the point-wise multiplication on each color channel of the image. See Fig. 2 for the illustration of the CutMix operation.

To assign labels to the mixed image X_m, [49] employ a Mean Teacher (MT) [48] model to assign pseudo-labels to the target image. In particular, they feed the target image X_t to the teacher segmentation model to obtain its pseudo-label

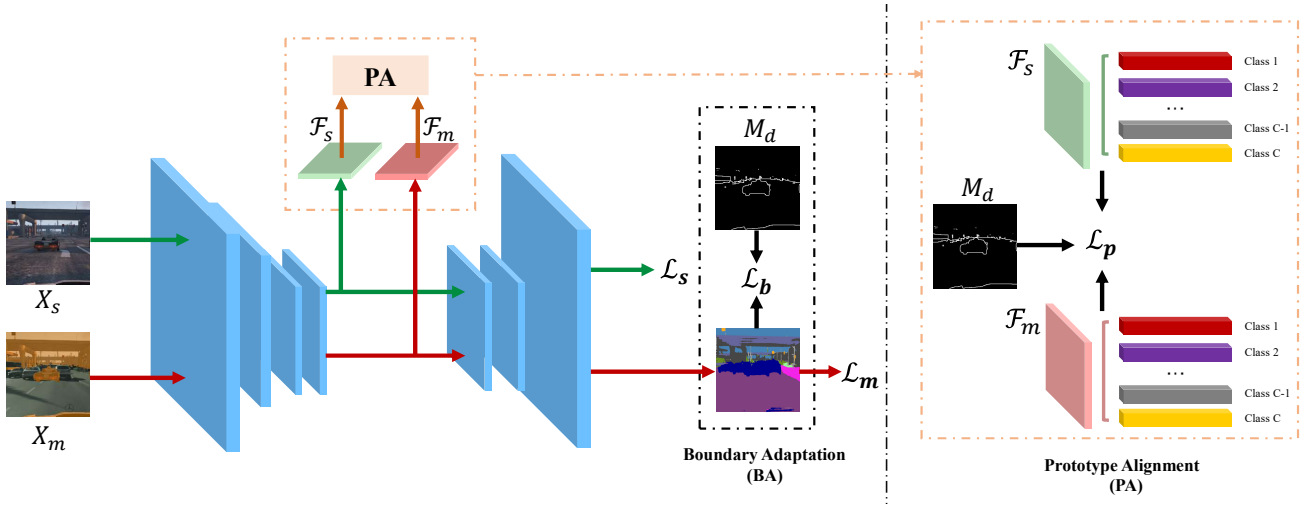


Figure 3. Overview of our proposed Boundary Adaptation and Prototype Alignment Network (BAPA-Net). The data flows of the source and mixed images are denoted by green and red lines respectively. The source image and the mixed image are used to optimize the semantic segmentation loss under the supervision of source ground truth and the mixed pseudo labels (*i.e.*, $\mathcal{L}_s$ and $\mathcal{L}_m$ in Eq. (9)). Our boundary adaptation module leverages the distance map M_d to reweight the cross-entropy loss for mixed image (*i.e.*, $\mathcal{L}_b$). The prototype alignment module reduces the domain mismatch by minimizing the distance of class prototypes between source image X_s and mixed images X_m with boundary removal (*i.e.*, $\mathcal{L}_p$). We train our BAPA-Net in an end-to-end manner.

$\hat{Y}_t$, then the labels for the mixed image X_m can be obtained using the same cut-paste operator:

$$Y_m = M_{mask} \odot Y_s + (1 - M_{mask}) \odot \hat{Y}_t, \tag{2}$$

After obtaining (X_m, Y_m), [49] uses both the original source images and the mixed images to train the semantic segmentation model with the cross-entropy loss. Please refer to [49] for the details.

We employ CutMix for boundary adaptation to produce the so-called *domain-mixed boundary samples*. The procedure is as follows: We first randomly paste some source objects to target images, and the pixels next to the domain boundaries of the pasted objects are from different domains and often from different classes. When extracting features using convolutional neural networks, the receptive fields of those pixels may cover parts of the adjacent source and target objects, making the extracted features contain mixed information from both domains. We believe those domain-mixed features are important for cross-domain semantic segmentation. Therefore, we pay special attention by giving higher weights to optimize those boundary samples during model training, and gradually the domain-mixed boundary samples will become domain-invariant and help alleviate the cross-domain issue.

Boundary enhancement loss. With the generated domain-mixed boundary samples, we are ready to enhance the model discriminability by enforcing the segmentation to focus on those samples. Since we feed the mixed image X_m as a whole to train the model, we thus calculate a weight map M_b to assign higher weights to boundary samples in the segmentation loss. Specifically, we first calculate a distance map to describe the distance of each pixel to its nearest cut-paste boundary. Let us denote $\mathcal{X}_b$ as all the pixels exactly located at the cut-paste boundary. Given a pixel located at coordinate (h, w) of X_m, denoted as $X_m^{(h,w)}$, its distance map value can be calculated as:

$$M_d^{(h,w)} = \min_{x \in \mathcal{X}_b} d(X_m^{(h,w)}, x), \tag{3}$$

where $d(\cdot, \cdot)$ is the Euclidean distance of the coordinates of two pixels.

Then, the boundary weight map M_b can be obtained as:

$$M_b = \left(1 - \frac{M_d}{\max(M_d)}\right) \odot \mathbb{1}\left[M_d < \lambda_d\right], \tag{4}$$

where $\max(M_d)$ denotes the maximum value of M_d, and $\mathbb{1}[\cdot]$ is an indicator function. In other words, we consider only pixels with distances smaller than λ_d, and samples with smaller distances will gain higher weights. Thus, we can calculate the boundary enhancement loss $\mathcal{L}_b$ as:

$$\mathcal{L}_{\text{b}} = \frac{1}{HW} \sum^{H,W} M_b \odot \mathcal{L}_{ce}(X_m^{(h,w)}, \hat{Y}_m^{(h,w)}), \tag{5}$$

where the L_{ce} is the standard cross-entropy loss.

3.2. Prototype Alignment

We have employed domain-invariant samples to enhance the discriminability of the segmentation model. However, the source and mixed images may still exist a mismatch in

the feature space. Therefore we propose to reduce the distribution mismatch with prototype alignment.

We follow the strategy in recent prototype alignment works [46, 55, 57]. In particular, they calculate class prototypes using source and target domain images and then reduce the distance between the corresponding classes of source and target domains. For the unlabeled target domain images, the pseudo-labels predicted by the segmentation model are used to compute the target prototypes.

In our scenario, we use the mixed images as a substitute for the target images for prototype alignment. The reasons are two folds. First, aligning prototypes of source images and mixed images are able to reduce the distribution mismatch of source and target domains. We are aware that the prototypes calculated from mixed images would be different from those from target images. In fact, they are more like a kind of intermediate domain prototypes. However, when prototypes of source images and mixed images are well aligned, it immediately implies the prototypes of source images and target images are aligned, too. Second, since we have used mixed images for enhancing our model, it is expected to make more confident predictions on the mixed images than the target images. Using mixed images as a substitute for the target images would also produce high-quality prototypes for alignment.

However, the boundary pixels in the mixed images are difficult to be correctly predicted. As a result, using features of all pixels may introduce noise to the prototypes, which possibly harms the prototype alignment. Therefore, we propose to improve the prototype alignment by excluding the boundary examples.

To obtain the target domain prototypes, we follow the same strategy as in [57] except that we exclude the boundary examples that may introduce noise to the target class prototypes. We denote by $\mathbf{f}_m$, $\tilde{y}_m$ as the feature vector and the predicted label of a pixel from a mixed image. To ensure the robustness of the prototypes, we also need to exclude the boundary examples. Then the set of validate features for the c-th class as $\mathcal{F}_m^c = \{\mathbf{f}_m | \tilde{y}_m = c \text{ and } M_d > \lambda_d\}$, where M_d is its distance to the boundary calculated with Eq. (3), and λ_d is the threshold for filtering boundary examples in Eq. (4). Then, the prototype of the c-th class for the mixed image can be calculated as:

$$\mathbf{p}_m^c = \frac{1}{|\mathcal{F}_m^c|} \sum_{\mathbf{f}_m \in \mathcal{F}_m^c} \mathbf{f}_m. \tag{6}$$

The prototypes for the source domain can be similarly calculated. Due to the ground truth labels are available for source domain, we do not exclude the boundary. Concretely, we average the features belonging to the correct predicted region for one class as its prototype in a training batch. Given any source pixel from the source images

Algorithm 1 Framework of BAPA-Net.

Input: Source domain dataset $\mathcal{D}_S$, target domain dataset $\mathcal{D}_T$, the CutMix mask M_{mask}, student network $G_{std} = C_{std} \circ E_{std}$ where C_{std} is the segmentation classifier and E_{std} is the feature extractor, teacher network G_{tea} whose weights are updated by using the EMA (exponential moving average) of G_{std}, and maximum number iteration N.

Output: The final student model G_{std}

```
Initialize network parameters for G_std with MSCOCO pre-trained weights, and make the parameters of G_tea same as G_std
while N ≠ 0 do
    X_s, Y_s ~ D_S
    X_t ~ D_T
    Ŷ_t ← G_tea(X_t)
    X_m, Y_m ← using Eq. (1) (2)
    M_b ← using Eq. (3) (4) to calculate boundary map
    F_s ← E_std(X_s), F_m ← E_std(X_m)
    p_m^c, p_s^c ← using Eq. (6) (7) to calculate prototype for target and source domains
    L ← using Eq. (9)
    Compute ∇L by backpropagation and update parameters of G_std
    G_tea ← EMA (G_tea, G_std)
    N ← N − 1
end while
return G_std
```

in a training batch, let us denote by $\mathbf{f}_s$, $\tilde{y}_s$, y_s as its feature vector, the predicted label and the ground truth label, respectively. We then can define set of the correctly predict features for the c-th class as $\mathcal{F}_s^c = \{\mathbf{f}_s | \tilde{y}_s = c \text{ and } y_s = c\}$, then the prototype of the c-th class can be calculated as:

$$\mathbf{p}_s^c = \frac{1}{|\mathcal{F}_s^c|} \sum_{\mathbf{f}_s \in \mathcal{F}_s^c} \mathbf{f}_s. \tag{7}$$

To align the prototypes of two domains, we follow [57] to maintain a prototype bank for source images and then align the prototypes of mixed images to the prototypes in the bank. The prototype bank consists of a fixed number of prototypes for each class, which are updated at every mini-batch in a first-in-first-out manner. Let us denote $\mathbf{p}_{s,i}^c$ as the i-th prototype for the c-th class in the bank. we minimize the ℓ_1 distance between each prototype from the mixed images with the closest prototypes of the same class in the bank, which can be formulated as:

$$\mathcal{L}_p = \sum_{c=1}^{C} \min_i \left\| \mathbf{p}_m^c - \mathbf{p}_{s,i}^c \right\|_1. \tag{8}$$

3.3. Overall Model

Following [49], both source images and mixed images are used to train the segmentation network based on the MT model. Moreover, we additionally optimize the boundary enhancement loss in Eq. (5) and the prototype alignment loss in Eq. (8) to improve the discrimination and domain-invariant abilities of the segmentation model. The overall objective of our BAPA-Net can be written as:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_m + \lambda_b \mathcal{L}_b + \lambda_p \mathcal{L}_p \tag{9}$$

where $\mathcal{L}_s$ and $\mathcal{L}_m$ are the cross-entropy loss for the source and mixed images, respectively, and λ_b and λ_p are two tradeoff parameters. The implementation of BAPA-Net is presented in Algorithm 1.

4. Experiments

4.1. Experimental Setup

Following the common experimental setup for cross-domain semantic segmentation [50, 37, 38, 51, 52, 61, 66, 64], we evaluate our method on two synthetic-to-real scenarios, namely GTA5 to Cityscapes and SYNTHIA to Cityscapes. In both experiments, Cityscapes is regarded as the target domain, while GTA5 and SYNTHIA are regarded as the source domain, respectively.

- **Cityscapes** [7] is a popular semantic segmentation benchmark dataset for autonomous driving. The dataset is labeled with 19 classes. It consists of 2,975 images in the training set and 500 images in the validation set. We use the training set without labels as the unlabeled target domain and the 500 images in the validation set for evaluation.
- **GTA5** [43] is a large-scale synthetic dataset consisting of 24,966 images rendered from a computer game called Grand Theft Auto V (GTA5). Each image has the size of 1,912 $\times$ 1,052 which are annotated with the same 19 classes in Cityscapes.
- **SYNTHIA** [44] is also a synthetic dataset for semantic segmentation task. And its subset named SYNTHIA-RAND-CITYSCAPES includes a collection of 9,400 photo-realistic images, which has 16 common classes with Cityscapes.

Evaluation metric. We use Intersection over Union (IoU) as the evaluation metric. For each class, we compute it by the formula $IoU = \frac{TP}{TP+FP+FN}$, in which TP, FP and FN are defined in the confusion matrix as the numbers of true positive, false positive and false negative pixels, respectively. In addition, we compute the mean of IoUs (*i.e.*, mIoU) over all classes.

Implementation details. In our experiments, we follow the widely used implementation protocol in previous works [52, 19, 30, 31, 66, 64, 59, 57, 58, 54, 49, 41], using DeepLab-v2 [3] as our base segmentation model and ResNet101 [21] as our backbone. The backbone is pre-trained on ImageNet [10] and MSCOCO [32]. The Stochastic Gradient Descent (SGD) with Nesterov acceleration is used as the optimizer, and the initial learning rate is set to 2.5×10^{-4}, which is then decreased using polynomial decay with exponent 0.9 [3]. The weight decay and momentum are set to 5×10^{-4} and 0.9, respectively. We resize source images to 760$\times$1,280 and target images to 512$\times$1,024, then we extract random crops of size 512$\times$512 in both source and target images. When calculating the prototype, we use the output of last layer in ResNet101. The batch size is set to 2, *i.e*, 2 source images and 2 target images in a mini-batch, and we obtain the final model after 250k iterations. We empirically set $\lambda_d = 4$, $\lambda_b = 4$ and $\lambda_p = 0.005$ for our experiments. We implement our method using PyTorch on an Nvidia GeForce RTX 2080Ti.

4.2. Comparisons with the State-of-the-arts

The previous cross-domain semantic segmentation methods are mainly divided into three categories: 1) methods based on adversarial training including AdvEnt [52], BDL [30] and FADA [54], 2) methods based on self-training including PyCDA [31], CRST [66], R-MRNet [64], UDADT [57] and IAST [41], and 3) methods based on data augmentation including DLOW [19], FDA [59] and DACS [49]. Table 1 and 2 show the comparisons with the state-of-the-arts domain adaptation methods for semantic segmentation of the three types of methods respectively on the GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes setting. All the models are based on DeepLab V2 and ResNet101 as their backbone, except the PyCDA [31] which is based on PSPNet [63]. For the GTA5 $\rightarrow$ Cityscapes, our proposed approach achieves 57.4%, which outperforms existing state-of-the-arts methods with significant margins of $5.3\% \sim 15.1\%$ mIoU. For the task from SYNTHIA $\rightarrow$ Cityscapes, we report the mIoU on 16 classes and 13 classes (excluding "wall", "fence", "pole"). Our method improves the performance about $3.5\% \sim 12.1\%$ and $4.2\% \sim 13.2\%$ on 16 classes and 13 classes, respectively. And we achieve 53.3% and 61.2% respectively, both of them outperforming the corresponding results of baseline methods with large margins. This clearly validates the effectiveness of our method.

We also present qualitative examples of the segmentation results of our methods in Fig 4. The results from the "Source Only" and the original CutMix (DACS) are included for comparison. Both of our BAPA-Net and DACS exhibit superior segmentation results compared with the "Source Only" baseline and our BAPA-Net achieves even better performance than the DACS. We attribute this to the usage of BA and PA module for improving the discrimi-

Table 1. The mIoUs (in %) for GTA5 → Cityscapes. The ResNet-101 is used as the backbone network.

Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bike	mIoU
Source	63.3	15.7	59.4	8.6	15.2	18.3	26.9	15.0	80.5	15.3	73.0	51.0	17.8	59.7	28.2	33.1	3.5	23.2	16.7	32.9
AdvEnt [52]	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
DLOW [19]	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
BDL [30]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
PyCDA [31]	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
CRST [66]	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
R-MRNet [64]	90.4	31.2	85.1	36.9	25.6	37.5	**48.8**	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
FDA [59]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
UDADT [57]	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
LDRDA [58]	90.8	41.4	84.7	35.1	27.5	31.2	38.0	32.8	85.6	42.1	84.9	59.6	34.4	85.0	42.8	52.7	3.4	30.9	38.1	49.5
FADA [54]	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
DACS [49]	89.9	39.7	87.9	30.7	39.5	**38.5**	46.4	52.8	**88.0**	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
IAST [41]	93.8	57.8	85.1	**39.5**	26.7	26.2	43.1	34.7	84.9	32.9	88.0	62.6	29.0	87.3	39.2	49.6	23.2	34.7	39.6	51.5
Ours	**94.4**	**61.0**	**88.0**	26.8	**39.9**	38.3	46.1	**55.3**	87.8	**46.1**	**89.4**	**68.8**	**40.0**	**90.2**	**60.4**	**59.0**	0.00	**45.1**	**54.2**	**57.4**

Table 2. The mIoUs (in %) for SYNTHIA → Cityscapes. mIoU* denotes the mean IoU over 13 classes excluding those marked with *. Classes not evaluated are replaced by '-'. The ResNet-101 is used as the backbone network.

Method	road	sidewalk	building	wall*	fence*	pole*	light	sign	vegetation	sky	person	rider	car	bus	motorcycle	bike	mIoU*	mIoU
Source	36.3	14.6	68.8	9.2	0.2	24.4	5.6	9.1	69.0	79.4	52.5	11.3	49.8	9.5	11.0	20.7	33.7	29.5
AdvEnt [52]	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	48.0	41.2
BDL [30]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	51.4	-
PyCDA [31]	75.5	30.9	83.3	20.8	0.7	32.7	27.3	33.5	84.7	85.0	64.1	25.4	85.0	45.2	21.2	32.0	53.3	46.7
CRST [66]	67.7	32.2	73.9	10.7	1.6	**37.4**	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	50.1	43.8
R-MRNet [64]	87.6	41.9	83.1	14.7	1.7	36.2	**31.3**	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	**53.1**	54.9	47.9
FDA [59]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	**38.4**	51.1	52.5	-
UDADT [57]	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	-
LDRDA [58]	85.1	44.5	81.0	-	-	-	16.4	15.2	80.1	84.8	59.4	31.9	73.2	41.0	32.6	44.7	53.1	-
FADA [54]	84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	84.0	53.5	22.6	85.4	43.7	26.8	27.8	52.5	45.2
DACS [49]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	**90.8**	**67.6**	**38.3**	82.9	38.9	28.5	47.6	54.8	48.3
IAST [41]	81.9	41.5	83.3	17.7	**4.6**	32.3	30.9	28.8	83.4	85.0	65.5	30.8	86.5	38.2	33.1	52.7	57.0	49.8
Ours	**91.7**	**53.8**	**83.9**	**22.4**	0.8	34.9	30.5	**42.8**	**86.6**	88.2	66.0	34.1	**86.6**	**51.3**	29.4	50.5	**61.2**	**53.3**

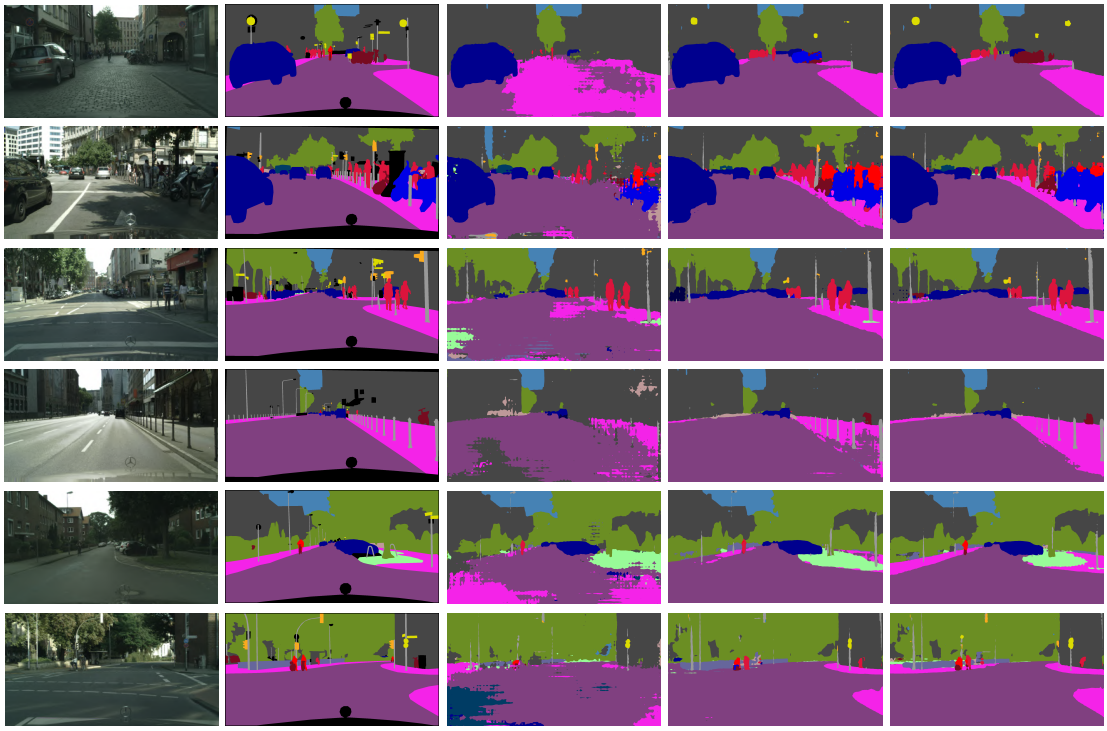
native and domain-invariant abilities of the segmentation model. For example, by using the boundary enhancement loss, our BAPA-Net generally predicts the boundary more precisely than DACS (see predicted the bikes and cars in the 1-st and 2-nd rows). Moreover, our BAPA-Net also does better at distinguishing road and sidewalk than DACS (see the 4-th, 5-th and 6-th rows), possibly because the prototype alignment helps to reduce the domain distribution mismatch.

4.3. Ablation Studies

In this section, we conduct the ablation experiments in the setting of GTA5 → Cityscapes.

Effects of different components. We validate the individual effects of our Boundary Adaptation (BA) and Prototype Alignment (PA) modules. The results are summarized in Table 3. We include the original CutMix model as a baseline for comparison. We conduct the ablation studies by removing different components. As shown in Table 3, despite baseline CutMix already achieves quite competitive performance, our proposed boundary adaptation and prototype alignment modules still gain large improvements, reaching 56.4% (w/o PA) and 55.8% (w/o BA) mIoU, respectively. By integrating both modules, our final BAPA-Net achieves an improvement of 5.3% over the baseline. These large improvements clearly validate that both of our modules play an important role in cross-domain semantic segmentation.

The impact of boundary removal in PA. In Section 3.2, we propose to calculate prototype that better reflects the class information for mixed images by excluding the features of boundary samples. Here we conduct an additional experiment for comparison by not removing these boundary samples, *i.e*, all pixel features are used for producing the prototypes for the mixed images. For a clear compari-



(a) Target Image (b) Ground Truth (c) Source Only (d) DACS (e) Ours

Figure 4. Qualitative segmentation results on the GTA5 → Cityscapes domain adaptation task. We present (a) Target Image, (b) Ground Truth, (c) Source Only, (d) DACS [49], (e) Ours.

Table 3. Ablation Studies on effects of different components

Model	mIoU	Δ
Baseline[49]	52.1	
BAPA-Net w/o BA	55.7	3.6↑
BAPA-Net w/o PA	56.4	4.3↑
BAPA-Net	57.4	5.3↑

Table 4. The impact of removing boundary samples in prototype alignment

	mIoU	Δ
w/o boundary removal	52.6	
with boundary removal	55.7	3.1↑

son, we consider the prototype alignment module only and do not use the Boundary Adaptation (BA) module in experiments. From Table 4, we observe that the PA module without using boundary removal leads to a performance drop of 3.1% than the proposed PA module. This clearly verifies our motivation that the boundary samples could mislead the prototype calculation in the mixed images.

5. Conclusion

In this paper, we address the problem of cross-domain semantic segmentation. We reveal a critical finding that previous works often neglect the importance of object boundaries while paying much attention to the overall segmentation results of whole objects. We empirically find that if we treat the object boundary properly, the segmentation performance can be considerably improved. Based on the observation, we present a novel method called Boundary Adaptation and Prototype Alignment Network (BAPA-Net), where we tackle the cross-domain semantic segmentation problem from two aspects. On the one hand, we employ the newly developed boundary adaptation strategy to focus more on the domain-mixed boundary samples, which are constructed based on CutMix and contain information from both the source and target domains. On the other hand, we design a prototype alignment module to reduce domain mismatch by minimizing the distance between class prototypes of the two domains, where boundary samples are ignored here to avoid domain confusion during the prototype calculation. Experiments on GTA5→Cityscapes and SYNTHIA→Cityscapes clearly validate the effectiveness of our BAPA-Net.

Acknowledgement: This work is partially supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400 and China Postdoctoral Science Foundation (NO. 2019TQ0051).

References

[1] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[2] Francesco Caliva, Claudia Iriondo, Alejandro Morales Martinez, Sharmila Majumdar, and Valentina Pedoia. Distance map loss penalty term for semantic segmentation. In *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.

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

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

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

[8] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[9] Bharath Bhushan Damodaran, Benjamin Kellenberger, Remi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

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

[11] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, June 2021.

[12] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[13] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] Lixin Duan, Ivor W. Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012.

[15] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[17] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[18] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.

[19] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

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

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

[23] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

[24] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, 2020.

[25] Minsu Kim, Sunghun Joung, Seungryong Kim, Jungin Park, Ig-Jae Kim, and K. Sohn. Cross-domain grouping and alignment for domain adaptive semantic segmentation. In *AAAI*, 2021.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q.

Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[27] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, 2020.

[28] Xiangtai Li, Xia Li, Li Zhang, Cheng Guangliang, Jianping Shi, Zhouchen Lin, Yunhai Tong, and Shaohua Tan. Improving semantic segmentation via decoupled body and edge supervision. In *ECCV*, 2020.

[29] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[30] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. 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.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 97–105, 2015.

[35] Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[36] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[37] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

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

[39] Fengmao Lv, Tao Liang, Xiang Chen, and Guosheng Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[40] Fengmao Lv, Guosheng Lin, Peng Liu, Guowu Yang, Sinno Jialin Pan, and Lixin Duan. Weakly-supervised cross-domain road scene segmentation via multi-level curriculum adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020.

[41] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *The European Conference on Computer Vision (ECCV)*, 2020.

[42] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[43] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

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

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[47] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[48] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[49] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1379–1389, January 2021.

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

[51] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[52] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[53] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. DADA: Depth-aware domain adaptation in semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

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

[55] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PANet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[56] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. RANet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[57] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[58] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Xinliang Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, 2020.

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

[60] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

[62] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[64] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

[65] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[66] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.