Context-Aware Mixup for Domain Adaptive Semantic Segmentation

Anonymous ICCV submission

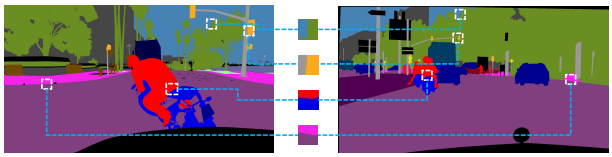
Paper ID 1516

Abstract

Unsupervised domain adaptation (UDA) aims to adapt a model of the labeled source domain to an unlabeled target domain. Although the domain shifts may exist in various dimensions such as appearance, textures, etc, the contextual dependency, which is generally shared across different domains, is neglected by recent methods. In this paper, we utilize this important clue as explicit prior knowledge and propose end-to-end Context-Aware Mixup (CAMix) for domain adaptive semantic segmentation. Firstly, we design a contextual mask generation strategy by leveraging accumulated spatial distributions and contextual relationships. The generated contextual mask is critical in this work and will guide the domain mixup. In addition, we define the significance mask to indicate where the pixels are credible. To alleviate the over-alignment (e.g., early performance degradation), the source and target significance masks are mixed based on the contextual mask into the mixed significance mask, and we introduce a significance-reweighted consistency loss on it. Experimental results show that the proposed method outperforms the state-of-the-art methods by a large margin on two widely-used domain adaptation benchmarks, i.e., GTAV → Cityscapes and SYNTHIA → Cityscapes. [1]

1. Introduction

Semantic segmentation aims to assign a semantic label to each pixel for a given image. Over the past few years, researchers have made great efforts to explore a variety of CNN methods trained on a large-scale segmentation dataset to tackle this problem [1, 2]. However, building such a large annotated dataset is both cost-expensive and time-consuming due to the process of annotating pixel-wise labels [9]. A natural idea to overcome this bottleneck is using synthetic data to supervise the segmentation model instead of real data [34, 35]. However, the existing domain gap between the synthetic images and real images often leads to a significant performance drop when the learned source models are directly applied to the unlabelled target data.



(a) source domain **(b) target domain**

Figure 1. Previous domain adaptation methods neglect the shared context dependency across different domains and could result in severe negative transfer and training instability. We observe that exploiting contexts as explicit prior knowledge is essential when adapting from the source domain to the target domain.

To address this issue, various unsupervised domain adaptation (UDA) techniques have been proposed to reduce the domain gap in pixel level [15, 52, 22, 18], feature level [24, 7, 58, 57] and output level [40, 25, 44, 5]. Among them, the most common practices are based on adversarial learning [40, 25, 44, 5], self-training [58, 57, 22, 31], consistency regularization [8, 48, 39], entropy minimization [41, 3] etc.

Previous works mainly focused on utilizing common prior knowledge, *e.g.,* appearances, scales, textures, weather, etc., to narrow down the domain gap. Nevertheless, context dependency across different domains has been very sparsely exploited so far in UDA, and how to transfer such cross-domain context still remains under-explored. As shown in Figure 1, we observe that the source and target images usually share similar semantic contexts, *e.g.,* rider is over the bicycle or motorcycle, sidewalk is beside the road, and such context knowledge is crucial particularly when adapting from the source domain to the target domain. The lack of context will lead to severe negative transfer, *e.g.,* early performance degradation during the adaptation process. In addition, most state-of-the-art approaches cannot be trained end-to-end. They heavily depend on the adversarial learning, image-to-image translation or pseudo labeling, and most of them need to fine-tune the models in many offline stages.

In this paper, we attempt to identify context dependency across domains as explicit prior domain knowledge when

[1] Our source code will be released.

adapting from the source domain to the target domain. We propose context-aware domain mixup (**CAMix**) to explore and transfer cross-domain contexts for domain adaptation. Our whole framework is fully end-to-end. The proposed CAMix consists of two key components: contextual mask generation (CMG) and significance-reweighted consistency loss (SRC).

To be specific, the CMG firstly generates a contextual mask by selectively leveraging the accumulated spatial distribution of the source domain and the contextual relationship of the target domain. This mask is critical in our work and will guide the domain mixup. Guided by it, context-aware domain mixup is performed in three different levels, *i.e.,* input level, output level and significance mask level. Notice that the significance mask is a mask that we define to indicate where the pixels are credible. This contextual mask respectively mixes the input images, the labels and the corresponding significance-masks to narrow down the domain gap.

In addition, we introduce a SRC loss on the significance mask level to alleviate the over-alignment, *e.g.,* early performance degradation, during the adaptation process. In particular, we calculate a significance mask with the help of the target predictive entropy and its dynamic threshold. Then, we mix the target and the source significance masks using the context knowledge as supervisory signals, and utilize the mixed significance mask to reweigh the consistency loss.

To sum up, we propose a *context-aware mixup* architecture for domain adaptive semantic segmentation, which is a fully end-to-end framework. Our contributions are summarized as follows.

- We present a *contextual mask generation* strategy, which leverages the spatial distribution of the source domain and the contextual relationship of the target domain. It acts as prior knowledge for guiding the context-aware domain mixup on three different levels.
- We introduce a *significance-reweighted consistency loss*, which alleviates the adverse impacts of the adaptation procedure, *e.g.,* early performance degradation and training instability, under the guidance of context.
- Extensive experimental results show that we outperform state-of-the-art methods by a large margin on two challenging UDA benchmarks. We achieve 55.2% mIoU in GTAV [34] $\rightarrow$ Cityscapes [9], and 59.7% mIoU in SYNTHIA [35] $\rightarrow$ Cityscapes [9], respectively.

2. Related Work

The current mainstream approaches for cross-domain semantic segmentation include adversarial learning [25, 40, 50, 5], consistency regularization [8, 48, 39] and self-training [58, 57]. As our work is mostly relevant to the latter two categories, we mainly focus on reviewing them.

Domain mixup: Mixup has been well-studied in other communities to improve the robustness of models., *e.g.,* semi-supervised learning [11, 12], and point cloud classification [54, 4]. A few works [47, 46, 26] studied cross-domain mixup in UDA. Nevertheless, these methods work well on simple and small classification datasets (*e.g.* MNIST [20] and SVHN [28]), but can hardly be applied to more challenging tasks, *e.g.,* domain adaptive semantic segmentation. DACS [39] is designed for segmentation, while little attention has been paid to exploiting contexts as prior knowledge to mitigate the domain gaps.

Consistency regularization: The key idea of consistency regularization is that the target prediction of the student model and that of the teacher model should be invariant under different perturbations. The teacher model is an exponential moving average (EMA) of the student model, and then the teacher model could transfer the learned knowledge to the student. Consistency regularization typically appears in Semi-supervised Learning (SSL) [38] and is recently applied to UDA recently [8, 33]. For simplicity, we choose [38] as a base framework to realize end-to-end learning.

Self-training: Self training [58, 57] aims to generate pseudo labels for the unlabeled target domain, and then fine-tuned the segmentation model on the pseudo labels iteratively in an offline way. Mei et al. [27] concentrated on the quality of pseudo labels and designed an instance adaptive self-training. Li et al. presented a self-supervised learning [22], which alternately trained the image translation model and the self-supervised segmentation adaptation model. In addition, CBST-BNN [13] and ESL [36] both leveraged predictive entropy rather than the maximum softmax predicted probabilities to refine the pseudo labels during the offline self-training. *Our method differs from these methods in several aspects*. Firstly, in contrast to previous offline self-training that generates pseudo labels and fine-tunes the segmentation model iteratively in many stages, our approach can be trained end-to-end in an online manner. Secondly, instead of using a probability-based mask in common self-training, *e.g.,* [58, 57], we calculate an entropy-guided mask with a novel significance-reweighted loss. Thirdly, different from [13, 36] to refine the pseudo labels, our significance mask is calculated based on the prior knowledge of context information.

Uncertainty estimation: The idea of exploiting model prediction uncertainty has been utilized in domain adaptation for classification, *e.g.,* Bayesian classifier [45] and Bayesian discriminator [19]. These methods always require an extra discriminator in adversarial training, and can work well on simple and small classification datasets. *Our method differs from these methods in several aspects.* At first, we tackle the

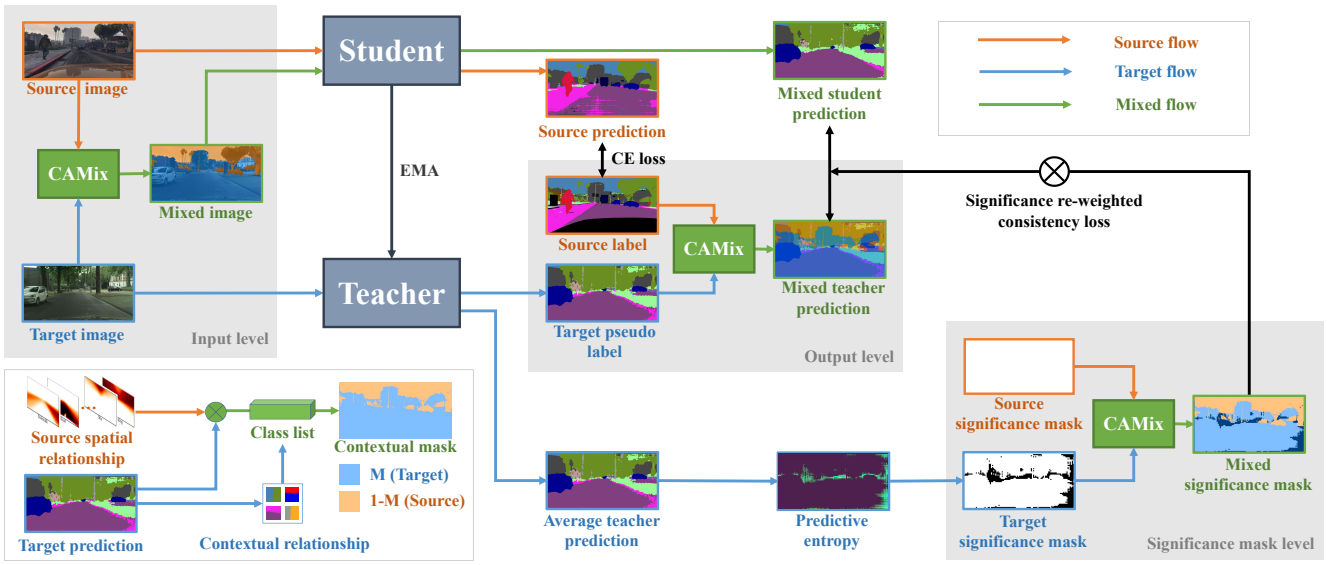


Figure 2. Overview of the proposed architecture. Firstly, we generate a contextual mask (CMG) by leveraging the spatial distribution of the source domain and the contextual relationship of the target domain. Guided by this mask M, we perform context-aware mixup (CAMix) in three levels, *i.e.,* input level, output level and significance mask level. Provided the context knowledge, we design a significance re-weighted consistency (SRC) loss to ease the over-alignment between the mixed student and teacher prediction.

more challenging task of semantic segmentation rather than image classification, where the uncertainty of dense pixel-wise predictions instead of image-wise prediction needs to be decreased. Secondly, we avoid using adversarial adaptation in uncertainty estimation which tends to be unstable and inaccurate. Thirdly, in comparison with the aforementioned approaches, we design significance mask level domain mixup between the target significance mask and the source mask, which enables a more informative entropy-guided mask during the domain mixup.

3. Methodology

Following the UDA protocols [40, 41, 58], we have access to the source images $X_S \in S$ with their corresponding labels Y_S. For the target domain T, only unlabeled images $X_T \in T$ are available. Unlike existing UDA methods that overlook the shared context knowledge across domains, we propose a novel context-aware domain mixup (CAMix) to exploit and transfer such cross-domain contexts.

Figure 2 shows the overview of our proposed architecture. Firstly, we present a contextual mask generation (CMG) strategy for mining the prior spatial distribution of the source domain and the contextual relationships of the target domain, thus generating a mask M. Guided by this mask, we perform an efficient CAMix on three levels, *i.e.,* input level, output level, and significance mask level (a mask we define to indicate where the pixels are credible.). In particular, the teacher model $f_{\theta'}$ is an exponential moving average (EMA) of the student model f_θ. In other words, the proposed CAMix uses the labeled source samples and unlabelled target samples to synthesize the mixed images, the mixed pseudo labels (Section 3.2), and the corresponding mixed significance masks (Section 3.3). We introduce a significance-reweighted consistency loss (SRC) on the significance mask (SigMask) level to alleviate the over-alignment during the online adaptation procedure.

3.1. Contextual Mask Generation

Intuitively, the source and the target domain share similar context dependency between domains. With this in mind, we identify two kinds of semantic contexts as explicit prior domain knowledge for guiding the domain adaptation procedure. The former is prior spatial contexts of the source domain, and the latter is contextual relationships of the categories in the target domain.

The scenes often have their intrinsic spatial structures, *e.g.,* sky tends to appear on the top of the image while roads are more likely to appear on the bottom. It is intuitive to explore the spatial relationships of the source domain. Thus, we generate a spatial prior matrix Q by counting the class frequencies in the source domain and we treat it as prior knowledge to regularize the target prediction: $\hat{f}_{\theta'} \leftarrow Q \odot f_{\theta'}(X_T)$, where $f_{\theta'}(T)$ is the target prediction of the teacher model.

To exploit the contextual relationship, *e.g.,* the traffic sign should be beside the pole, our core idea is to find the semantic-related categories of the current class presented in the image. In other words, these classes that have contextual relationships to each other can be treated as a meta-class, and then we copy them together from the target images and

paste them onto the the source images, which prevents certain semantic categories hanging on an inappropriate context.

Specifically, we first get the spatially-modulated pseudo label: $\tilde{Y}_T \leftarrow \arg\max_{c'} \hat{f}_{\theta'}(i,j,c')$. Next, we randomly select half of the classes present in the argmax prediction $\tilde{Y}_T$, namely c. After that, we judge whether each category $k \in c$ presented in $\tilde{Y}_T$ is in the meta-class list m or not. The meta-class list involves several groups of heuristic meta-classes, *e.g.,* pole, traffic sign, traffic light, bicycle, motorcycle and rider, etc, and this list is shared in all experiments. If $k \in c$, we append the semantic-related classes $\tilde{k}$ of class k to the current list c.

A binary contextual mask M is generated by setting the pixels from the final class list c to value 1 in M, and all others to value 0.

$$M(i,j) = \begin{cases} 1, \text{ if } \tilde{Y}_T(i,j) \in c \\ 0, \text{ otherwise} \end{cases}, \tag{1}$$

where $i \in h, j \in w$. We iterate each spatial location to generate the mask. *This mask M is then utilized as prior knowledge* to mix the images in the input level, the labels in output level (Section 3.2), and the significance mask on the significance mask level (Section 3.3) between the source domain and the target domain.

3.2. Input-level and Output-level Domain Mixup

In the *input level*, the image X_S and X_T sampled from the source domain and target domain are synthesized into X_M:

$$X_M = M \odot X_T + (1 - M) \odot X_S, \tag{2}$$

where $\odot$ denotes element-wise multiplication.

The weights $\Phi_t^{'}$ of the teacher model at training step t are updated by the student's weights Φ_t with a smoothing coefficient $\alpha \in [0,1]$, which can be formulated as follows:

$$\Phi_t^{'} = \alpha \cdot \Phi_{t-1}^{'} + (1 - \alpha) \cdot \Phi_t, \tag{3}$$

where α is the EMA decay that controls the updating rate.

Regarding the *output level*, the label of source domain Y_S and the pseudo label of target domain $\hat{Y}_T = f_{\theta'}(X_T)$ are mixed as:

$$Y_M = M \odot \hat{Y}_T + (1 - M) \odot Y_S. \tag{4}$$

Different from [39, 29], we mix the images and the corresponding labels in a target-to-source direction rather than the source-to-target direction. In other words, we copy some categories from the target domain and paste them onto the source domain, where we can add our consideration of both spatial relationship and contextual relationship in such a direction.

3.3. Significance-mask Level Domain Mixup

In the significance-mask (SigMask) level domain mixup, we decrease the uncertainties of the mixed teacher prediction with the guidance of contextual mask M as additional supervisory signals. As a result, we are able to alleviate the adverse impact, *e.g.,* training instability and early performance degradation, and transfer more reasonable knowledge from the teacher to the student.

Stochastic forward passes. In particular, we repeat each target image for N copies and inject a random Gaussian noise for each target sample copy. Then, given a set of pixel-wise predicted class scores $\{\boldsymbol{P}_i^{(h,w,c)}(x_t)\}_{i=1}^N$ of target samples, we can get the mean of the predictive probability $\hat{P}_c$ of the c-th class:

$$\hat{P}_c = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{P}_i^{(h,w,c)}(X_T). \tag{5}$$

Note that we do not use any dropout layers during stochastic forward passes. The predictive entropy ζ is calculated as:

$$\zeta^{(h,w)} = -\sum_{c=1}^{C} \hat{P}_c \cdot log(\hat{P}_c), \tag{6}$$

where all volumes of pixel-wise entropy forms a set $K = \{\zeta\}_{i=1}^N$.

Dynamic threshold. A dynamic threshold H is then determined by the predictive entropy rather than the softmax probabilities to filter out the unreliable pixel-wise predictions:

$$H = \beta + (1 - \beta) \cdot e^{\gamma(1 - t/t_{max})^2} \cdot K_{sup}, \tag{7}$$

where t denotes the current training step and t_{max} is the maximum training step. K_{sup} means the upper-bound of the volumes' self-information, which is denoted as: $K_{sup} = sup\{\zeta\}_{i=1}^N$. We use the same β and γ by default in all experiments.

Significance mask. We denote the SigMask $U_T = I(\zeta < H)$, where I is an indicator function. Note that although the predictive entropy $\zeta^{(h,w)}$ is similar to ADVENT [41], we do not perform entropy minimization at all, and our SigMask U_T is calculated from Eq. (5) to Eq. (7) in a completely different way, with the help of target predictive entropy ζ and its dynamic threshold H.

Given the contextual mask M as additional supervisory signals, we perform **SigMask level** domain mixup. The significance mask of the source domain U_S and the target domain U_T are mixed into U_M:

$$U_M = M \odot U_T + (1 - M) \odot U_S, \tag{8}$$

where U_S is a tensor full of 1, because the source labels are provided without uncertainties. And these certain areas do

not need to reweigh the consistency loss. Only the uncertain areas in the target U_T which is below the dynamic threshold H, are set to 0 to reweigh the consistency loss.

Significance-reweighted consistency loss. To encourage the teacher model to transfer more credible knowledge to the student model, we define a SRC loss with the guidance of U_M:

$$\mathcal{L}_{con}\left(f_{\theta'}, f_{\theta}\right) = \frac{\sum_j \left(U_M \cdot CE(f_\theta(X_M), Y_M)\right)}{\sum_j U_M}, \quad (9)$$

where $f_{\theta'}$ and f_θ are the teacher model and the student model, respectively. CE is the abbreviation of the cross-entropy loss. The pixel-wise SigMask U_M is used to reweigh the consistency loss in a weighted averaging manner.

Algorithm 1: Context-Aware Mixup

Input: student model f_θ, teacher model $f_{\theta'}$, source domain D_S, target domain D_T, total iterations N.
Output: teacher model $f_{\theta'}$.

Initialize network parameters θ randomly. ;
for *i=1* **to** *N* **do**
  $X_S, Y_S \sim \mathcal{D}_S$;
  $X_T \sim \mathcal{D}_T$;
  $\hat{Y}_T \leftarrow f_{\theta'}(X_T)$;
  $X_M \leftarrow$ Input-level mixup by Eq.(2);
  $\hat{Y}_S \leftarrow f_\theta\left(X_S\right), \hat{Y}_M \leftarrow f_\theta\left(X_M\right)$;
  $Y_M \leftarrow$ Output-level mixup by Eq.(4);
  $U_T \leftarrow$ Target SigMask by Eq.(5)∼Eq.(7) ;
  $U_M \leftarrow$ SigMask-level mixup by Eq.(8);
  $\mathcal{L}_{total} \leftarrow$ Total loss by Eq.(11);
  Compute $\nabla_\theta \mathcal{L}_{total}$ by backpropagation;
  Perform stochastic gradient descent on θ;
end
return $f_{\theta'}$;

3.4. End-to-end Training

Segmentation loss. The segmentation loss L_{seg} is a cross-entropy loss for optimizing the images from the source domain:

$$\mathcal{L}_{seg} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} Y_S^{(h,w,c)} log(P_S^{(h,w,c)}), \quad (10)$$

where Y_S is the ground truth for source images and $P_S = f_\theta(X_S)^{(h,w,c)})$ is the segmentation output of source images.

Total loss. During training, all models on three different levels are jointly trained in an end-to-end manner. The whole framework is optimized by integrating all the aforementioned loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{con}\mathcal{L}_{con}, \quad (11)$$

where λ_{con} is the weight of consistency loss. The algorithm of CAMix for the whole training process is illustrated in Algorithm 1.

4. Experiments

Following common UDA protocols [16, 40], we treat the labeled synthetic dataset, *i.e.,* GTAV [34] and SYNTHA [35], as the source domain, and the unlabeled real dataset *i.e.,* Cityscapes [9] as the target domain.

4.1. Datasets

Cityscapes [9] is a dataset focused on autonomous driving, which consists of 2,975 images in the training set and 500 images in the validation set. The images have a fixed spatial resolution of 2048 × 1024 pixels. Following common practice, we trained the model on the unlabelled training set and report our results on the validation set.

GTAV [34] is a synthetic dataset including 24,966 photo-realistic images rendered by the gaming engine Grand Theft Auto V (GTAV). The semantic categories are compatible between the two datasets. We used all the 19 official training classes in our experiments.

SYNTHIA [35] is another synthetic dataset composed of 9,400 annotated synthetic images with the resolution of 1280 × 960. Like GTAV, it has semantically compatible annotations with Cityscapes. Following the prior works [7, 55, 6], we use the SYNTHIA-RAND-CITYSCAPES subset [35] as our training set.

4.2. Implementation Details

In our implementation, we employ DeepLab-v2 [1] with ResNet 101 backbone [14]. The backbone is pre-trained on ImageNet [10] and MSCOCO [23]. For the DeepLab-v2 network, we use Adam as the optimizer. The initial learning rate is 2.5×10^{-4} which is then decreased using polynomial decay with an exponent of 0.9. The weight decay is 5×10^{-5} and the momentum is 0.9. Following the common UDA protocol [22, 25], when the source domain is GTAV, we resize all images to 1280×720; when the source domain is SYNTHIA, we resize all images to 1280×760. Then, both the source and target images are randomly cropped to 512×512. We use the same data augmentation as DACS [39], i.e., color jittering and Gaussian blurring. In our SigMask-level CAMix, we perform $N = 8$ times of stochastic forward passes. Following the previous consistency regularization works, we use the same adaptive schedule as CutMix [11] and DACS [39] for the consistency weight λ_{con}. Our method is implemented in Pytorch on a single NVIDIA Tesla V100. *More details can be found in the supplementary material.*

Table 1. Comparison results (mIoU) from GTAV to Cityscapes.

Method	Venue	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motocycle	bike	**mIoU**
Source Only	-	63.3	15.7	59.4	8.6	15.2	18.3	26.9	15.0	80.5	15.3	73.0	51.0	17.7	59.7	28.2	33.1	3.5	23.2	16.7	32.9
BDL [22]	CVPR'19	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
APODA [50]	AAAI'20	85.6	32.8	79.0	29.5	25.5	26.8	34.6	19.9	83.7	40.6	77.9	59.2	28.3	84.6	34.6	49.2	8.0	32.6	39.6	45.9
IntraDA [30]	CVPR'20	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
SIM [44]	CVPR'20	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
LTIR [18]	CVPR'20	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA [52]	CVPR'20	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
PCEDA [51]	CVPR'20	91.0	49.2	85.6	37.2	29.7	33.7	38.1	39.2	85.4	35.4	85.1	61.1	32.8	84.1	45.6	46.9	0.0	34.2	44.5	50.5
LSE [37]	ECCV'20	90.2	40.0	83.5	31.9	26.4	32.6	38.7	37.5	81.0	34.2	84.6	61.6	33.4	82.5	32.8	45.9	6.7	29.1	30.6	47.5
WLabel [32]	ECCV'20	91.6	47.4	84.0	30.4	28.3	31.4	37.4	35.4	83.9	38.3	83.9	61.2	28.2	83.7	28.8	41.3	8.8	24.7	46.4	48.2
CrCDA [17]	ECCV'20	92.4	55.3	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6
FADA [43]	ECCV'20	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
LDR [49]	ECCV'20	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
CCM [21]	ECCV'20	93.5	57.6	84.6	39.3	24.1	25.2	35.0	17.3	85.0	40.6	86.5	58.7	28.7	85.8	49.0	56.4	5.4	31.9	43.2	49.9
DAST [53]	AAAI'21	92.2	49.0	84.3	36.5	28.9	33.9	38.8	28.4	84.9	41.6	83.2	60.0	28.7	87.2	45.0	45.3	7.4	33.8	32.8	49.6
Ours	-	**93.3**	**58.2**	**86.5**	36.8	**31.5**	**36.4**	35.0	**43.5**	**87.2**	**44.6**	**88.1**	**65.0**	24.7	**89.7**	**46.9**	**56.8**	**27.5**	**41.1**	**56.0**	**55.2**

4.3. Comparison with the State-of-the-Arts

Table 1 and Table 3 present the comparison results with the state-of-the-arts on two challenging tasks: "GTAV $\rightarrow$ Cityscapes" and "SYNTHIA $\rightarrow$ Cityscapes". Our proposed method significantly outperforms the state-of-the-art techniques by $5\% \sim 10\%$ on GTAV $\rightarrow$ Cityscapes and $6\% \sim 12\%$ on SYNTHIA $\rightarrow$ Cityscapes. Also, it is superior to the non-adaptive baselines by around 22% and 30% on two benchmarks, respectively.

Most of the state-of-the-art approaches perform the adversarial learning, *e.g.,* APODA [50], IntraDA [30], WLabel [32], MRNet [56], FADA [43] and DADA [42], and they need to carefully tune the optimization procedure for min-max problems through a domain discriminator. However, such domain discriminators tend to be unstable and inaccurate. Instead, our method does not require to maintain an extra discriminator during the domain adaptation process, and we outperform these approaches by $6\% \sim 10\%$ in mIoU.

In contrast to the offline self-training methods that need to fine-tune the models in many rounds, *e.g.,* CRST [57], LSE [37], CCM [21] and TPLD [31], our whole framework can be trained in a fully end-to-end manner. Benefiting from the online consistency regularization by our proposed components CMG and SRC, our approach significantly outperforms the self-training methods by around $5\% \sim 9\%$.

Compared to the methods which require an image-to-image (I2I) translation or style transfer algorithm to filter out the domain-specific texture or style information, *e.g.,* BDL [22], LDR [49], LTIR [18], FDA [52] and PCEDA [51], our context-aware domain mixup does not require any style/spectral transfer algorithms or deep neural networks for I2I translation. Our domain mixup algorithm is simple and works very well, and it surpasses the translation-based methods by around $5\% \sim 8\%$.

CrCDA [17] learned and enforced the prototypical local contextual-relations in the feature space, while the visual cues of context knowledge tend to be lost. Moreover, such an learning does not *explicitly* exploit the cross-domain contexts and cannot be trained end-to-end. In contrast, our CAMix explicitly explores the contexts in the image space rather than the feature space, and our architecture can be trained end-to-end. Our approach outperforms the CrCDA [17] by 6.6% and 9.7% in two benchmarks, respectively.

Taking a closer look at per-category performance in Table 1 and Table 3, our approach achieves the highest IoU on most categories, *e.g.,* motorcycle, bicycle, traffic sign, etc. This phenomenon reveals the effectiveness of CAMix among different classes during the adaptation process.

Table 2. Comparisons with existing domain mixup methods.

method	mIoU (%)	Gain (%)
Mean Teacher	43.1	–
+ CowMix [12]	48.3	+5.2
+ CutMix [11]	48.7	+5.6
+ DACS [39]	52.1	+9.0
+ iDACS [39]	51.5	+8.4
+ CAMix	**55.2**	**+12.1**

4.4. Comparison with the Other Domain Mixup

As shown in Table 2, we present the adaptation results of our method and the existing domain mixup algorithms on GTAV $\rightarrow$ Cityscapes. We choose the Mean Teacher architecture [38] as our baseline in this experiment. The existing domain mixup algorithms are implemented under the same settings. CowMix [12], CutMix [11] are proposed for semi-supervised learning (SSL), and we adapt them to the UDA task, which mixes the source domain image and the target domain image. Besides, we implement the existing cross-

Table 3. Comparison results (mIoU) from SYNTHIA to Cityscapes.

Method	Venue	road	sidewalk	building	light	sign	vegetation	sky	person	rider	car	bus	motocycle	bike	**mIoU$_{13}$**
Source Only	-	36.3	14.6	68.8	5.6	9.1	69.0	79.4	52.5	11.3	49.8	9.5	11.0	20.7	29.5
BDL [22]	CVPR'19	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
DADA [42]	ICCV'19	89.2	44.8	81.4	8.6	11.1	81.8	84.0	54.7	19.3	79.7	40.7	14.0	38.8	49.8
APODA [50]	AAAI'20	86.4	41.3	79.3	22.6	17.3	80.3	81.6	56.9	21.0	84.1	49.1	24.6	45.7	53.1
IntraDA [30]	CVPR'20	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
LTIR [18]	CVPR'20	92.6	53.2	79.2	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	49.3
SIM [44]	CVPR'20	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
FDA [52]	CVPR'20	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
LSE [37]	ECCV'20	82.9	43.1	78.1	9.1	14.4	77.0	83.5	58.1	25.9	71.9	38.0	29.4	31.2	49.4
CrCDA [17]	ECCV'20	86.2	44.9	79.5	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	50.0
WLabel [32]	ECCV'20	92.0	53.5	80.9	3.8	6.0	81.6	84.4	60.8	24.4	80.5	39.0	26.0	41.7	51.9
CCM [21]	ECCV'20	79.6	36.4	80.6	22.4	14.9	81.8	77.4	56.8	25.9	80.7	45.3	29.9	52.0	52.9
LDR [49]	ECCV'20	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
DAST [53]	AAAI'21	87.1	44.5	82.3	13.9	13.1	81.6	86.0	60.3	25.1	83.1	40.1	24.4	40.5	52.5
Ours	-	**91.8**	**54.9**	**83.6**	**23.0**	**29.0**	**83.8**	**87.1**	**65.0**	26.4	**85.5**	**55.1**	36.8	**54.1**	**59.7**

Table 4. Ablation study of each component in CAMix.

iDACS [39]	SP	CR	SRC	mIoU
√				51.5
√	√			53.1
√	√	√		54.1
√	√	√	√	55.2

Table 5. Ablation studuy of the SRC loss

Baseline	Mixup	$\mathcal{L}_{con}$	mIoU	△
iDACS [39]	CMG	SRC	55.2	-
	CMG	MSE	44.5	↓ 9.7
	CMG	CE	54.2	↓ 1.0

Table 6. Ablation study of each level in CAMix.

MT	SigMask	In-Out	mIoU (GTAV)	mIoU$_{13}$ (SYN)
√			43.1	45.9
√	√		44.6	47.1
√	√	√	55.2	59.7

domain mixup method, *e.g.,* DACS [39] and inverse DACS. The former DACS means using ClassMix to copy the source categories and paste them onto the target. Inverse DACS (iDACS) [39] uses a target-to-source direction.

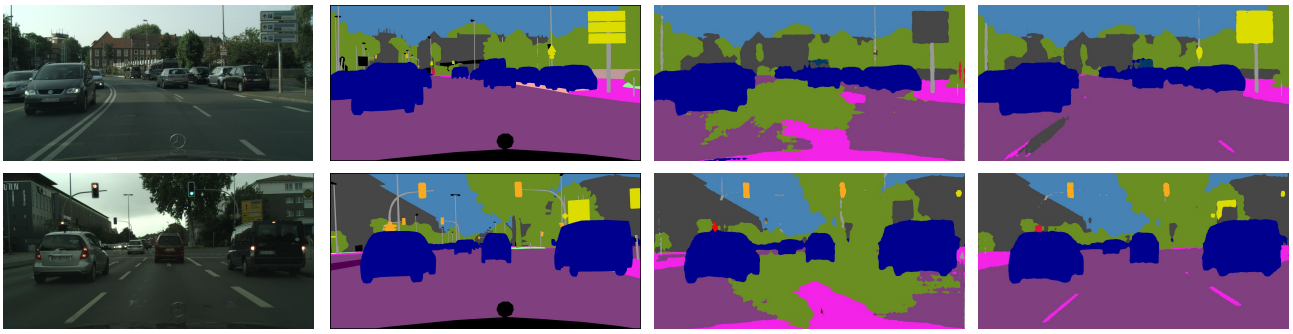
We analyze that using CowMix [12] results in the occurrence of partial objects in the mixed images, which are harder to learn in the training process. Besides, CutMix [11], DACS [39] and iDACS tend to result in severe label contamination and category confusion when generating the mixed results, thus leading to negative transfer. The main reason is that they neglect the context dependency as prior knowledge for facilitating the domain adaptation. The results shown in Table 2 demonstrate the superiority of our proposed CAMix to other domain mixup algorithms.

4.5. Ablation Studies

In this section, we study the effectiveness of each component (Table 4) and each level (Table 6) in our approach and investigate how they contribute to the final performance when adapting from the GTAV [34] to Cityscapes [9].

Effectiveness of CMG: The CMG strategy is a fundamental component of our framework, which is designed to capture the shared context dependency across domains for CAMix. *Spatial prior (SP)* and *contextual relationship (CR)* are two key components of CMG. The ablation studies of each component in CAMix are reported in Table 4. Compared to the baseline (iDACS) [39], SP and CR could successfully bring 1.6% and 2.6% of improvements, achieving 53.1% and 54.1% on the former two levels, respectively. By adding the SRC loss on the SigMask level, we can achieve an even higher performance of 55.2%.

Effectiveness of SRC: Table 5 shows the contribution of the SRC loss on the GTAV $\rightarrow$ Cityscapes benchmark. The full CAmix with all three levels and SRC loss achieves 55.2%. If we directly replace the SRC loss with a normal *mean square error (MSE)*, the result is even worse and only reaches 44.5%. Using the *cross-entropy (CE)* as the consistency loss boosts the mIoU to 54.2%, which is still 1.0%



(a) Input Image (b) Ground Truth (c) DACS (d) Ours

Figure 3. Qualitative segmentation results in the SYNTHIA → Cityscapes setup. The four columns plot (a) RGB input image, (b) ground-truth, (c) the predictions of DACS [39] and (d) the predictions of our CAMix.

worse than our SRC loss in Eq. (9). The main benefits of the SRC loss are reflected as follows. The SigMask-level domain mixup with the SRC loss could further decrease the uncertainty of the teacher model, and promote the teacher model to transfer reasonable knowledge to the student, thus improving the performance. As such, our approach tends to be more stable and effectively ease these negative impacts, *i.e.,* training instability and early performance degradation, during the adaptation process.

Effectiveness of different levels: Table 6 lists the impacts of different levels on the above two settings, *i.e.,* taking GTAV and SYNTHIA as the source domains, respectively. The Mean Teacher (MT) baseline achieves 43.1% and 45.9% on two benchmarks, respectively. By adding the SigMask-level domain mixup, our method respectively brings +1.5% and +1.2% improvements. By integrating the CAMix on three levels together, we finally achieve 55.2% and 59.7% mIoU, respectively.

4.6. Visualization

Qualitative segmentation results. Figure 3 visualizes some segmentation results on the task SYNTHIA → Cityscapes. As we can see from the figure, due to the lack of context dependency, DACS [39] incorrectly classifies some large categories *e.g.,* the road as sidewalk or terrain, and produces some false predictions on some sophisticated classes, *e.g.,* traffic sign. Our proposed method is capable of outputting high confidence predictions compared to the previous work.

Performance curve of adaptation. Figure 4 plots the performance curves to show the effectiveness of SRC loss when adapting from GTAV [34] to Cityscapes [9]. Previous methods, *e.g.,* Mean Teacher [8], neglect the context knowledge shared by different domains and perform a rough distribution matching, resulting in training instability and early performance degradation. Instead, we effectively ease these negative impacts and decrease the uncertainty of segmentation model, by introducing the SRC loss.

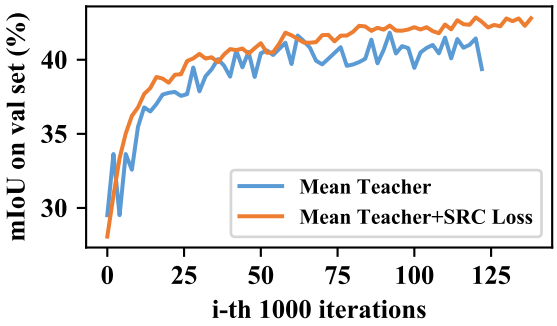


Figure 4. Comparison on adapting from GTA5 [34] dataset to Cityscapes [9] dataset. The blue line corresponds to the conventional consistency regularization strategy [8]. The orange line indicates the consistency-based adaptation with our SRC loss. Our method can ease the issue of training instability and early performance drop.

5. Conclusion

In this paper, we proposed a novel context-aware domain mixup (CAMix) framework for domain adaptive semantic segmentation. We present a contextual mask generation (CMG) strategy, which is critical for guiding the whole pipeline on three different levels, *i.e.,* input level, output level and significance mask level. Our approach can explicitly explore and transfer the shared context dependency across domains, thus narrowing down the domain gap. We also introduce a significance-reweighted consistency loss (SRC) to penalize the inconsistency between the mixed student prediction and the mixed teacher prediction, which effectively eases the adverse impacts of the adaptation, *e.g.,* training instability and early performance degradation. The extensive experiments with ablation studies demonstrate that our approach soundly outperforms the state-of-the-art methods in domain adaptive semantic segmentation.

References

[1] 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. *TPAMI*, 40(4):834–848, 2017. 1, 5

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[3] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019. 1

[4] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. In *ECCV*, 2020. 2

[5] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 1, 2

[6] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018. 5

[7] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017. 1, 5

[8] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *ICCV*, 2019. 1, 2, 8

[9] 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 *CVPR*, 2016. 1, 2, 5, 7, 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[11] Geoff French, Samuli Laine, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020. 2, 5, 6, 7

[12] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020. 2, 6, 7

[13] Ligong Han, Yang Zou, Ruijiang Gao, Lezi Wang, and Dimitris Metaxas. Unsupervised domain adaptation via calibrating uncertainties. In *CVPR Workshops*, 2019. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 1

[16] 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. 5

[17] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. *arXiv preprint arXiv:2007.02424*, 2020. 6, 7

[18] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *CVPR*, 2020. 1, 6, 7

[19] Vinod Kumar Kurmi, Shanu Kumar, and Vinay P Namboodiri. Attending to discriminative certainty for domain adaptation. In *CVPR*, 2019. 2

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[21] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, 2020. 6, 7

[22] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 1, 2, 5, 6, 7

[23] 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 *ECCV*, 2014. 5

[24] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019. 1

[25] 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 *CVPR*, 2019. 1, 2, 5

[26] Xudong Mao, Yun Ma, Zhenguo Yang, Yangbin Chen, and Qing Li. Virtual mixup training for unsupervised domain adaptation. *arXiv preprint arXiv:1905.04215*, 2019. 2

[27] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020. 2

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 2

[29] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. *arXiv preprint arXiv:2007.07936*, 2020. 4

[30] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, 2020. 6, 7

[31] Inkyu Shin Sanghyun Woo Fei Pan and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *ECCV*, 2020. 1, 6

[32] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schulter, Amit K Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. *arXiv preprint arXiv:2007.15176*, 2020. 6, 7

[33] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. 2

[34] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 1, 2, 5, 7, 8

[35] 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 *CVPR*, 2016. 1, 2, 5

[36] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Esl: Entropy-guided self-supervised learning for domain adaptation in semantic segmentation. In *CVPR Workshops*, 2020. 2

[37] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. *arXiv preprint arXiv:2007.14449*, 2020. 6, 7

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2, 6

[39] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. *arXiv preprint arXiv:2007.08702*, 2020. 1, 2, 4, 5, 6, 7, 8

[40] 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 *CVPR*, 2018. 1, 2, 3, 5

[41] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 1, 3, 4

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 6, 7

[43] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, 2020. 6

[44] 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 *CVPR*, 2020. 1, 6, 7

[45] Jun Wen, Nenggan Zheng, Junsong Yuan, Zhefeng Gong, and Changyou Chen. Bayesian uncertainty matching for unsupervised domain adaptation. *arXiv preprint arXiv:1906.09693*, 2019. 2

[46] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *ECCV*, 2020. 2

[47] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020. 2

[48] Yonghao Xu, Bo Du, Lefei Zhang, Qian Zhang, Guoli Wang, and Liangpei Zhang. Self-ensembling attention networks: Addressing domain shift for semantic segmentation. In *AAAI*, 2019. 1, 2

[49] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. *arXiv preprint arXiv:2003.04614*, 2020. 6, 7

[50] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *AAAI*, 2020. 2, 6, 7

[51] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *CVPR*, 2020. 6

[52] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 1, 6, 7

[53] Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *AAAI*, 2021. 6, 7

[54] Jinlai Zhang, Lvjie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujing Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. *arXiv preprint arXiv:2101.01461*, 2021. 2

[55] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 5

[56] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *arXiv preprint arXiv:1912.11164*, 2019. 6

[57] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 1, 2, 6

[58] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 1, 2, 3