ML-BPM: Multi-teacher Learning with Bidirectional Photometric Mixing for Open Compound Domain Adaptation in Semantic Segmentation

Fei Pan[1], Sungsu Hur[1], Seokju Lee[2], Junsik Kim[3], and In So Kweon[1]

[1] KAIST, South Korea. {feipan, sshuh1215, iskweon77}@kaist.ac.kr
[2] KENTECH, South Korea. slee@kentech.ac.kr
[3] Harvard University, USA. mibastro@gmail.com

Abstract. Open compound domain adaptation (OCDA) considers the target domain as the compound of multiple unknown homogeneous subdomains. The goal of OCDA is to minimize the domain gap between the labeled source domain and the unlabeled compound target domain, which benefits the model generalization to the unseen domains. Current OCDA for semantic segmentation methods adopt manual domain separation and employ a single model to simultaneously adapt to all the target subdomains. However, adapting to a target subdomain might hinder the model from adapting to other dissimilar target subdomains, which leads to limited performance. In this work, we introduce a multi-teacher framework with bidirectional photometric mixing to separately adapt to every target subdomain. First, we present an automatic domain separation to find the optimal number of subdomains. On this basis, we propose a multi-teacher framework in which each teacher model uses bidirectional photometric mixing to adapt to one target subdomain. Furthermore, we conduct an adaptive distillation to learn a student model and apply consistency regularization to improve the student generalization. Experimental results on benchmark datasets show the efficacy of the proposed approach for both the compound domain and the open domains against existing state-of-the-art approaches.

Keywords: Domain Adaptation, Open Compound Domain Adaptation, Semantic Segmentation, Multi-teacher Distillation

1 Introduction

Semantic segmentation is a fundamental task in finding applications to many problems, including robotics [34], autonomous driving [35], and medical diagnosis [16]. Recently, deep learning-based semantic segmentation approaches [12,36,35] have achieved remarkable progress. However, their effectiveness and generalization ability require a large amount of pixel-wised annotated data which are expensive to collect. To reduce the cost of data collection and annotation, numerous synthetic datasets have been proposed [21,22]. However, the models trained on

synthetic data tend to poorly generalize to real images. To cope with this issue, unsupervised domain adaptation (UDA) methods [26,28,37,17,25,33] have proposed to align the domain gap between the source and the target domain. Despite the efficacy of UDA techniques, most of these works rely on the strong assumption that the target data is composed of a single homogeneous domain. This assumption is often violated in real-world scenarios. As an illustration in autonomous driving, the target data will likely be composed of various subdomains such as night, snow, rain, etc. Therefore, directly applying the current UDA approaches to these target data might deliver limited performance. This paper focuses on the challenging problem of open compound domain adaptation (OCDA) in semantic segmentation where the target domain is unlabeled and contains multiple homogeneous subdomains. The goal of OCDA is to adapt a model to a compound target domain and to further enhance the model generalization to the unseen domains.

To perform OCDA, Liu *et al.* [13] propose an easy-to-hard curriculum learning strategy, where samples closer to the source domain will be chosen first for adaptation. However, it does not fully take advantage of the subdomain boundaries information in the compound target domain. To explicitly consider this information, current OCDA works [8,19] propose to separate the target compound domain into multiple subdomains based on image style information. Existing works use a manual domain separation method; they also employ a single model to simultaneously adapt to all the target subdomain. However, adapting to a target subdomain might hinder the model from adapting to other dissimilar target subdomains, which leads to limited performance. We propose a multi-teacher framework with bidirectional photometric mixing for open compound domain adaptation in semantic segmentation to tackle this issue. First, we propose automatic domain separation to find the optimal number of subdomains and split the target compound domain. Then, we present a multi-teacher framework in which each teacher model uses bidirectional photometric mixing to adapt to one target subdomain. On this basis, we conduct adaptive distillation to learn a student model and apply a fast and short online updating using consistency regularization to improve the student's generalization to the open domains. We evaluate our approach on the benchmark datasets. The proposed approach outperforms all the existing state-of-the-art OCDA techniques and the latest UDA techniques for domain adaptation and domain generalization task.

The Contribution of This Work. (1) we propose automatic domain separation to find the optimal number of target subdomains; (2) we present a multi-teacher framework with bidirectional photometric mixing to reduce the domain gaps between the source domain and every target subdomain separately; (3) we further conduct an adaptive distillation to learn a student model and apply consistency regularization to improve the student generalization to the open domains.

2 Related Work

Unsupervised Domain Adaptation. Unsupervised domain adaptation (UDA) techniques are used to reduce the expensive cost of pixel-wise labeling tasks like semantic segmentation. In UDA, adversarial learning is used actively to align input-level style using image translation, feature distribution, or structured output [27,10,28,17,29]. Alternatively, self-training approaches [2,33,25,37] have also recently demonstrated compelling performance in this context. While these works have shown significant improvement, adopting those works directly for practical usage shows limitations due to its restricted setting dealing with only single source and single target. Despite the improvement provided by UDA techniques, their applicability to real scenarios remains restricted by the implicit assumption that the target data contains images from a single distribution.

Domain Generalization. The purpose of domain generalization (DG) is to train a model – solely using source domain data – such that it can perform reliable predictions on unseen domain. While DG is an essential problem, a few works have attempted to address this problem in the task of semantic segmentation. DG for semantic segmentation shows two main streams: augmentation-based and network-based approaches. The augmentation-based approaches [30,11] propose to significantly augment the training data via an additional style dataset to learn domain-invariant representation. The network-based approaches [18,4] attempt to modify the structure of the network to minimize domain-specific information (such as colors or styles) such that the resulting model mainly focuses on the content-specific information. Even though DG for semantic segmentation has achieve obvious progress, their performance is inevitably lower than several UDA methods due to the absence of the target images, which is capable of providing abundant domain-specific information.

Open Compound Domain Adaptation. Liu *et al.* [13] firstly suggests Open Compound Domain Adaptation (OCDA) that handles unlabeled compound heterogeneous target domain and unseen open domain. While Liu *et al.* [13] propose a curriculum learning strategy, it fails to consider the specific information of each target subdomain. Current OCDA works [8,19] propose to separate the compound target domain into multiple subdomains to handle the intra-domain gaps. Gong *et al.* [8] adopt domain-specific batch normalization for adaptation. Park *et al.* [19] utilize GAN-based image translation and adversarial training to exploit domain invariant features from multiple subdomains.

3 Generating Optimal Subdomains

3.1 Automatic Domain Separation

Our work assumes that the domain-specific property of images comes from their styles. Existing works adopt a predefined parameter to decide the number of subdomains, which might lead to a nonoptimal domain adaptation performance; furthermore, they rely on a pre-trained CNN-based encoder to extract the style information for the subdomain discovery. However, we propose an automatic

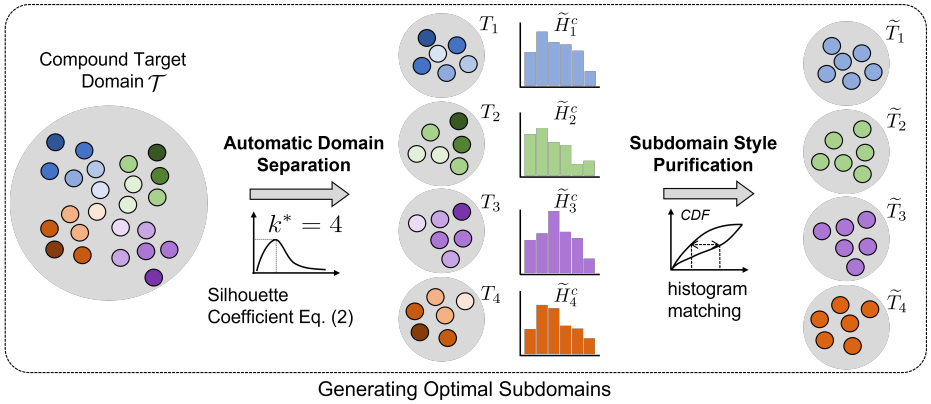

Fig. 1: The part of generating optimal subdomains consists of automatic domain separation (ADS) and subdomain style purification (SSP). In ADS, we adopt Silhouette Coefficient [23] to find the optimal number of subdomains k^*. In SSP, we calculate mean of histogram $\widetilde{H}_m^c$ for the m^{th} target subdomain T_m according to Equation 4, and the purified subdomain is denoted as $\widetilde{T}_m$.

domain separation (ADS) to effectively separate the target domain using the distribution of pixel values of the target images. The proposed ADS is capable of predicting the optimal number of subdomains without relying on any predefined parameters and extracting the image style information without relying on any pre-trained CNN models. We denote the source domain as $\mathcal{S}$, and the unlabeled compound target domain as $\mathcal{T}$. We also assume compound target domain contains k latent subdomains: $\{T_1, \ldots, T_k\}$, which lack of clear prior knowledge to distinguish themselves. The goal of ADS is to find the optimal number of subdomains k^* and separate $\mathcal{T}$ into several subdomains accordingly.

Current work [14] suggests a simple yet effective style translation method by matching the distribution of pixel values on LAB color space. Thus, we adopt LAB space into ADS to extract the style information of the target image. Given a target RGB image $x_t \in \mathcal{T}$ as input, we convert it into LAB color space $rgb2lab(x_t)$. The three channels in LAB color space are represented as l, a, and b. Then, we compute the histograms of the pixel values for all three channels in LAB color space: $H^l(x_t), H^a(x_t)$, and $H^b(x_t)$. The histograms are concatenated and represented as the style information of x_t. Let $s(x_t) = H^l(x_t) \frown H^a(x_t) \frown H^b(x_t)$ denote the concatenated histograms of x_t, and we take $s(x_t)$ as input to ADS for domain separation. However, most existing clustering algorithms require a hyperparameter to determine the number of clusters. Directly applying a naive clustering might lead to a nonoptimal adaptation performance. Thus, we propose to find the optimal number k^* of the subdomains using Silhouette Coefficient (SC) [23]. Suppose the target domain $\mathcal{T}$ is separated into k subdomains, $\{T_1, \ldots, T_k\}$. For each target image x_t, we denote $\gamma(x_t)$ as the average distance between x_t and all other target images in the target subdomain to which x_t

belongs. Additionally, we use $\delta(x_t)$ to represent the minimum average distance from x_t to all other target subdomains to which x_t does not belong. Let us assume x_t belongs to the m^{th} target subdomain T_m, then $\gamma(x_t)$ and $\delta(x_t)$ are written as

$$\begin{aligned}\gamma(x_t) &= \frac{\sum_{x_t{}' \in T_m, x_t{}' \neq x_t} L(s(x_t{}'), s(x_t))}{|T_m| - 1}, \\ \delta(x_t) &= \min_{T_n : 1 \leq n \leq k, n \neq m} \left\{ \frac{\sum_{x_t{}' \in T_n} L(s(x_t{}'), s(x_t))}{|T_n|} \right\},\end{aligned} \tag{1}$$

where $L(s(x_t{}'), s(x_t))$ represents the euclidean distance of $s(x_t{}')$ and $s(x_t)$, and $|T_m|$ is the number of the target images in T_m. The SC score for k number of the target subdomains is given by

$$SC(k) = \sum_{x_t \in \mathcal{T}} \frac{\delta(x_t) - \gamma(x_t)}{\max(\gamma(x_t), \delta(x_t))}. \tag{2}$$

Hence, the goal of the proposed ADS is to find k^* for

$$k^* = \arg\max_k SC(k). \tag{3}$$

3.2 Subdomain Style Purification

With the help fo automatic domain separation, the number of abnormal samples with different styles is small inside each target subdomain. Though these abnormal samples might be useful for the model's generalization, they could also lead to a negative transfer, which further hinders the model from learning domain invariant features in a specific subdomain. To cope with it, we propose to purify the style distribution of the target images inside each subdomain. We design a subdomain style purification (SSP) module to effectively make similar styles for the images within the same subdomain. Given the m^{th} target subdomain T_m, we adopt the histograms of LAB color space $\{(H^l(x_t), H^a(x_t), H^b(x_t)); \forall x_t \in T_m\}$ (mentioned in 3.1), and then we compute the mean of the histograms for all the three channels, represented by $\widetilde{H}^l_m, \widetilde{H}^a_m$, and $\widetilde{H}^b_m$, and this process is achieved by

$$\widetilde{H}^c_m = \frac{\sum_{x_t \in T_m} H^c(x_t)}{|T_m|}; \forall c \in \{l, a, b\}, \tag{4}$$

where $|T_m|$ represents the number of the target images in T_m. We take $\{\widetilde{H}^l_m, \widetilde{H}^a_m, \widetilde{H}^b_m\}$ as the standard style for T_m. For each target RGB image $x_t \in T_m$, we change the style of x_t to generate the RGB new image $\tilde{x}_t$ by the histogram matching [20] on $\widetilde{H}^l_m, \widetilde{H}^a_m$, and $\widetilde{H}^b_m$ on the LAB color space. The process of SSP is done for all the subdomains $\{T_1, \ldots, T_{k^*}\}$. We denote the purified subdomains after SSP as $\{\widetilde{T}_1, \ldots, \widetilde{T}_{k^*}\}$.

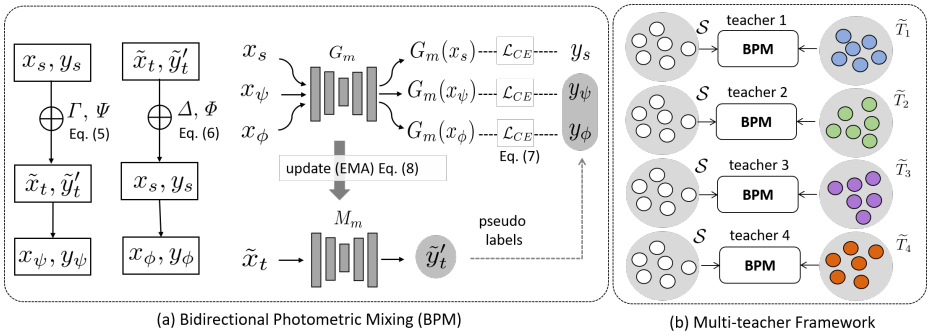

Fig. 2: (a) The architecture of the proposed bidirectional photometric mixing. (b) The diagram of the multi-teacher learning framework.

4 Multi-teacher Framework

4.1 Bidirectional Photometric Mixing

Through automatic domain separation and subdomain style purification (mentioned in 3.1 and 3.2), the compound domain $\mathcal{T}$ is automatically separated into multiple subdomains $\{\widetilde{T}_1, \ldots, \widetilde{T}_{k^*}\}$, where k^* represents the optimal number of the subdomains. Our next plan is to minimize the domain gap between the source domain and each target subdomain. A recent UDA work DACS [25] presents a mixing-based UDA technique for semantic segmentation. Inspired by DACS, we propose bidirectional photometric mixing (BPM) to minimize the domain gap between the source domain and each target subdomain separately. Compared with DACS, the proposed BPM adopts a photometric transform to decrease the style inconsistency of the mixed images to reduce the pixel-level domain gap. On this basis, BPM applies a bidirectional mixing scheme to provide a more robust regularization for training. The architecture of BPM is shown in Figure 2(a). The proposed BPM contains a domain adaptive segmentation network G_m and a momentum network M_m that improves the stability of pseudo labels. Let $(x_s, y_s) \in \mathcal{S}$ denote the source RGB image and its pixel-wise annotation map, $x_s \in \mathbb{R}^{H\times W\times 3}$, $y_s \in \mathbb{R}^{H\times W}$. And $(\tilde{x}_t) \in \widetilde{T}_m$ represent a purified target RGB image from the m^{th} purified subdomain $\widetilde{T}_m$, $\tilde{x}_t \in \mathbb{R}^{H\times W\times 3}$. Note that H and W represent the size of height and width. Our BPM applies the mixing in two directions: $\mathcal{S} \to \widetilde{T}_m$ and $\widetilde{T}_m \to \mathcal{S}$.

On the direction of mixing from $\mathcal{S} \to \widetilde{T}_m$, we choose ClassMix [15] because the source image x_s has the pixel-wise annotation map y_s. We first randomly select some classes from y_s. Then, we define $\Psi \in \{0,1\}^{H\times W}$ as a binary mask in which $\Psi(h,w) = 1$ when the pixel position (h,w) of x_s belongs to the selected classes, and $\Psi(h,w) = 0$ otherwise. While ClassMix suggests directly copying the corresponding pixels of selected classes of x_s onto $\tilde{x}_t$, the mixed image generated by ClassMix contains inconsistent style distribution which might hinder the adaptation performance. To cope with the limitation, the proposed BPM

applies photometric transform Γ on the selected source pixels to the style of target image before directly copying them onto it. Let $\Psi \odot x_s$ represent the selected source pixels by the mask Ψ, and $\odot$ is element-wise multiplication. We first calculate the histograms of selected source pixels in LAB color space, and match them with $\{\widetilde{H}_m^l, \widetilde{H}_m^a, \widetilde{H}_m^b\}$. The translated source pixels is represented as $\Gamma(\Psi \odot x_s)$. Then, we copy the translated source pixels onto $\tilde{x}_t$. We present some qualitative results in Figure 4. Note that no ground-truth annotation is available for $\tilde{x}_t$. Thus, we send the purified target image $\tilde{x}_t$ to the momentum network M_m to generate a stable prediction map $\tilde{y}'_t$ as the pseudo label. The mixing process on the direction of $\mathcal{S} \rightarrow \widetilde{T}_m$ by BPM is shown as

$$\begin{aligned} x_\psi &= \Gamma(\Psi \odot x_s) + (\mathbf{1} - \Psi) \odot \tilde{x}_t, \\ y_\psi &= \Psi \odot y_s + (\mathbf{1} - \Psi) \odot \tilde{y}'_t, \end{aligned} \tag{5}$$

where x_ψ is the generated mixed image, y_ψ is the corresponding mixed pseudo label, and $\Gamma(\cdot)$ is the photometric transform of the source selected pixels by histogram matching on LAB color space.

On the direction of mixing from $\widetilde{T}_m \rightarrow \mathcal{S}$, however, it is impossible to choose ClassMix since no ground-truth annotation is available for $\tilde{x}_t$. Inspired by CutMix [31], we generate another binary mask $\Phi \in \{0,1\}^{H \times W}$ by sampling rectangular bounding box (d_x, d_y, d_w, d_h) according to the uniform distribution; $d_x \sim U(0, W), d_y \sim U(0, H), d_w = W\sqrt{1-\eta}, d_h = H\sqrt{1-\eta}$, where $\eta \sim U(0,1)$, (H, W) are the height and width of the image. The binary mask Φ is formed by filling with 1 the pixel positions inside the bounding box, and filling with 0 other positions. With the help of Φ, we select the target pixels $\Phi \odot \tilde{x}_t$ and transform them into the source style. The transformed target pixel is represented by $\Delta(\Phi \odot \tilde{x}_t)$. Then we paste them onto the source image x_s. We present the mixing of $\widetilde{T}_m \rightarrow \mathcal{S}$ at

$$\begin{aligned} x_\phi &= \Delta(\Phi \odot \tilde{x}_t) + (\mathbf{1} - \Phi) \odot x_s, \\ y_\phi &= \Phi \odot \tilde{y}'_t + (\mathbf{1} - \Phi) \odot y_s, \end{aligned} \tag{6}$$

where x_ϕ is the other generated mixed image, y_ϕ is the corresponding mixed pseudo label, and $\Delta(\cdot)$ is the photometric transform of the target selected pixels by histogram matching on LAB color space.

we (x_ψ, y_ψ) (x_ϕ, y_ϕ) and (x_s, y_s) to train the segmentation network G_m and the momentum network M_m. We first optimize the parameters of G_m through

$$\begin{aligned} \mathcal{L}_{BGM}(\theta_m) = \sum_{\forall x_s \in S} \sum_{\forall \tilde{x}_t \in \widetilde{T}_m} \Big[& \mathcal{L}_{CE}\Big(G_m(x_s), y_s\Big) + \alpha \mathcal{L}_{CE}\Big(G_m(x_\psi), y_\psi\Big) \\ & + \beta \mathcal{L}_{CE}\Big(G_m(x_\phi), y_\phi\Big) \Big], \end{aligned} \tag{7}$$

where θ_m represent the parameters of G_m, $\mathcal{L}_{CE}$ is the cross-entropy loss for the predicted segmentation maps and the ground-truth or pseudo labels, α and β are the hyper-parameters to control the effect of the mixing of both the directions for the loss function. To help the momentum network M_m provide stable pseudo labels, we update the parameters of M_m, represented by θ'_m, using an exponential

moving average (EMA) with a momentum $\lambda \in [0, 1]$. After finishing the training iteration t, ${\theta'_m}$ is updated by

$$ {\theta'_m}^{t+1} = \lambda {\theta'_m}^{t} + (1 - \lambda)\theta_m. \tag{8} $$

4.2 Multi-teacher Adaptive Knowledge Distillation

We propose a multi-teacher framework followed by an adaptive knowledge distillation to align the domain gaps between the source domain and all the target subdomains. Given a purified subdomain $\widetilde{T}_m$, we adopt a BPM as a specific teacher model to minimize the domain gap between $\mathcal{S}$ and $\widetilde{T}_m$. And we train the proposed multi-teacher framework by minimizing the loss function $\mathcal{L}_{MT}$ on all the teacher models, *i.e.*,

$$ \mathcal{L}_{MT} = \sum_{m=1}^{k^*} \mathcal{L}_{BGM}(\theta_m), \tag{9} $$

where $\mathcal{L}_{BGM}(\theta_m)$ (defined in Equation 7) is the loss function of the segmentation network G_m in the m^{th} teacher model, and k^* is the optimal number of the subdomains. Moreover, We learn a segmentation network G_{sd} as the student network via an adaptive knowledge distillation from all the teacher networks $\{G_m : 1 \leq m \leq k^*\}$. Given a random target data from $x_t \in \mathcal{T}$, we send x_t to all the teachers model, and the student is to learn from a weighted average of the all teacher's predictions $O_w(x_t)$, based on the teacher's confidence score. We adopt the entropy of G_m's prediction map $G_m(x_t) \in \mathbb{R}^{H \times W \times C}$ as the confidence of the m^{th} teacher model, where C is the total number of classes we consider. Thus, the weight w_m for the m^{th} teacher and the average prediction $G_{out}(x_t)$ are formulated as

$$ \begin{aligned} w_m &= \frac{\sum_{h,w,c} G_m(x_t) \log \left[G_m(x_t)\right]}{\sum_{m'} \sum_{h,w,c} G_{m'}(x_t) \log \left[G_{m'}(x_t)\right]}, \\ G_{out}(x_t) &= \sum_{m=1}^{k^*} w_m G_m(x_t). \end{aligned} \tag{10} $$

On this basis, we optimize the student segmentation network G_{sd} with a distillation loss $\mathcal{L}_D$ defined by

$$ \mathcal{L}_D = \sum_{x_t \in \mathcal{T}} \mathcal{L}_{KL}\Big[G_{sd}(x_t) || G_{out}(x_t)\Big], \tag{11} $$

where $\mathcal{L}_{KL}$ is KL divergence loss function between the output of G_{sd} and G_{out}. The goal of the multi-teacher adaptive knowledge distillation is to achieve the optimal parameters ${\theta_{sd}}^*$ of the student segmentation network G_{sd} by

$$ {\theta_{sd}}^* = \min_{\theta_{sd}} \mathcal{L}_{MT} + \mathcal{L}_D. \tag{12} $$

Online Updating with Consistency Regularization. To evaluate the generalization of our approach, we directly evaluate our student network on the open domains as shown in Table 2a and Table 2b. Additionally, after finishing the compound domain adaptation training, we also provide a fast and short online updating for the student network using consistency regularization. This would further boost the generalization of the student network. Given an RGB image x_o from an open domain, we first match the style of x_o to other standard styles from the existing target subdomains. The standard styles are defined as the mean histograms $\{\widetilde{H}_m^l, \widetilde{H}_m^a, \widetilde{H}_m^b\}$ (defined in 3.2). The newly transformed images are $\{x_o^m; 1 \leq m \leq k^*\}$, where x_o^m is generated by matching x_o to the style of the m^{th} subdomain $\widetilde{T}_m$. Thus, we conduct an online updating for the student network G_{sd} by

$$\min_{\theta_{sd}} \sum_{m=1}^{k^*} \mathcal{L}_1\big(G_{sd}(x_o^m), G_{sd}(x_o)\big), \tag{13}$$

where $\mathcal{L}_1$ is the mean absolute loss. After the online updating, we test the student network with newly learnt parameters again on the open domains.

5 Experiments

5.1 Experimental Setup

Dataset. In this work, we adopt the synthetic datasets, including GTA5 [21] and SYNTHIA [22] as the source domains. GTA5 contains $24,966$ annotated images of $1,914 \times 1,052$ resolution. SYNTHIA consists of $9,400$ images with $1,280 \times 760$ resolution. Furthermore, we adopt C-Driving [13] as the compound target domains which contains real images of $1,280 \times 720$ resolution collected from different weather conditions. Following the settings of previous works [13,19,8], we use the $14,697$ rainy, snowy, cloudy images as the compound target domain and adopt 627 overcast images as the open domain. We also use ACDC [24] as another compound target domain and the evaluation results are shown in supplementary material. We further adopt Cityscapes [5], KITTI [1], and WildDash [32] as the open domains to evaluate the generalization ability of the proposed approach.

Implementation Details. We adopt DeepLab-V2 [3] with ResNet101 backbone [9] pre-trained on ImageNet [6]. All the images from target domain are rescaled into $1,280 \times 720$ and then randomly cropped into 640×360. The batch size is set up with 2 and the total number of training iterations is 2.5×10^5. We adopt stochastic gradient descent to optimize all the segmentation networks, with a weight decay of 5×10^{-4} and momentum of 0.9. The learning rate is set up with an initial value of 2.5×10^{-4} and decreased by polynomial decay with an exponent of 0.9. The momentum network has the same network architecture as the segmentation network. Existing mixing techniques contain CutMix [31], CowMix [7] and ClassMix [15]. We adopt ClassMix on the mixing direction of

Table 1: The performance comparison of mean IoU on the compound domain. Our approach is compared with the state-of-the-art UDA and OCDA approaches on (a) GTA5→C-Driving and (b) SYNTHIA→C-Driving benchmark dataset with ResNet-101 as the backbone. Note that mIoU^{11} represents the mean IoU of 11 classes, excluding the class with *.

(a) GTA5→C-Driving

Method	Type	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source	-	73.4	12.5	62.8	6.0	15.8	19.4	10.9	21.1	54.6	13.9	76.7	34.5	12.4	68.1	31.0	12.8	0.0	10.1	1.9	28.3
CDAS [13]	OCDA	79.1	9.4	67.2	12.3	15.0	20.1	14.8	23.8	65.0	22.9	82.6	40.4	7.2	73.0	27.1	18.3	0.0	16.1	1.5	31.4
CSFU [8]	OCDA	80.1	12.2	70.8	9.4	24.5	22.8	19.1	30.3	68.5	28.9	82.7	47.0	16.4	79.9	36.6	18.8	0.0	13.5	1.4	34.9
SAC [2]	UDA	81.5	23.8	72.0	10.3	27.8	23.0	18.2	34.1	70.3	27.9	87.8	45.0	16.9	77.6	38.5	19.8	0.0	14.0	2.7	36.4
DACS [25]	UDA	81.9	24.0	72.2	11.9	28.6	24.2	18.3	35.4	**71.8**	28.0	87.7	44.9	15.6	78.4	39.1	**24.9**	0.1	6.9	1.9	36.6
DHA[19]	OCDA	79.9	14.5	71.4	**13.1**	32.0	**27.1**	20.7	35.3	70.5	27.5	86.4	47.3	23.3	77.6	44.0	18.0	0.1	13.7	2.5	37.1
Ours	OCDA	**85.3**	**26.2**	**72.8**	10.6	**33.1**	26.9	**24.6**	**39.4**	70.8	**32.5**	**87.9**	**47.6**	**29.2**	**84.8**	**46.0**	22.8	**0.2**	**16.7**	**5.8**	**40.2**

(b) SYNTHIA→C-Driving

Method	Type	road	sidewalk	building	wall	fence	pole	light	sign*	veg	sky	person	rider*	car	bus*	mbike*	bike*	mIoU^{16}	mIoU^{11}
Source	-	33.9	11.9	42.5	1.5	0.0	14.7	0.0	1.3	56.8	76.5	13.3	7.4	57.8	12.5	2.1	1.6	20.9	28.1
CDAS [13]	OCDA	54.5	13.0	53.9	0.8	0.0	18.2	13.0	13.2	**60.0**	78.9	17.6	3.1	64.2	12.2	2.1	1.5	25.3	34.0
CSFU [8]	OCDA	69.6	12.2	50.9	1.3	0.0	16.7	12.1	13.6	56.2	75.8	20.0	4.8	68.2	14.1	0.9	1.2	26.1	34.8
SAC [2]	UDA	69.8	13.4	56.2	1.7	0.0	20.0	9.6	13.7	52.5	78.1	29.1	15.5	68.9	10.9	**3.2**	1.2	27.7	36.3
DACS [25]	UDA	62.1	**15.2**	48.8	0.3	0.0	19.7	10.3	9.6	57.8	**84.4**	35.2	18.9	67.8	16.0	2.2	1.7	28.1	36.5
DHA [19]	OCDA	67.5	2.5	54.6	0.2	0.0	**25.8**	13.4	**27.1**	58.0	83.9	36.0	6.1	71.6	**28.9**	2.2	1.8	29.9	37.6
Ours	OCDA	**73.4**	**15.2**	**57.1**	**1.8**	0.0	23.2	**13.5**	23.9	59.9	83.3	**40.3**	**22.3**	**72.2**	23.3	2.3	**2.2**	**32.1**	**40.0**

the source domain to the target domain, and we apply CutMix on the mixing direction of the target domain to the source domain. Both α and β are set up with 1 in the experiments. To increase the robustness of the segmentation model, we adopt data augmentations, including flipping, color jittering, and Gaussian blurring on the mixed images.

5.2 Results

To demonstrate the efficacy of our approach, we conduct experiments on the benchmark datasets of GTA5→C-Driving and SYNTHIA→C-Driving. We first compare our approach with the existing state-of-the-art OCDA approaches: CDAS [13], DHA [19], and CSFU [8]. Furthermore, we compare the proposed approach with the current state-of-the-art UDA approaches SAC [2] and DACS [25].

Compound Domain Adaptation. We first compare the performance of our approach with existing state-of-the-art OCDA and UDA approaches on GTA5 →C-Driving, shown in Table 1a. All the results are generated on the validation set of C-Driving. Training only with the source data leads to 28.3% of mean IoU over the 19 classes. As the first work in OCDA, CDAS achieves 31.4% on the mean IoU of all the classes. CSFU generates 34.9% of mean IoU, and DHA produces 37.1% of mean IoU. This is because both CSFU and DHA adopt the subdomain separation step and GAN framework, and DHA uses a more effective

Table 2: The comparison of mean IoU on the open domains. The domain generalization (DG) model is trained only with the source domain. All the models are tested on the validation set of C-Driving Open (O), cityscapes (C), KITTI (K), and wildDash (W). We also present the scores of our approach without online updating (w/o Updating) and with online updating (w/ Updating).

(a) GTA5 as the source domain.

GTA5						
Method	Type	O	C	K	W	Avg
CSFU [8]	OCDA	38.9	38.6	37.9	29.1	36.1
DACS [25]	UDA	39.7	37.0	40.2	30.7	36.9
RobustNet [4]	DG	38.1	38.3	40.5	30.8	37.0
DHC [19]	OCDA	39.4	38.8	40.1	30.9	37.5
Ours (w/o Updating)	OCDA	41.8	40.9	44.0	32.9	40.0
Ours (w/ Updating)	OCDA	42.5	41.7	44.3	34.6	40.8

(b) SYNTHIA as the source domain.

SYNTHIA						
Method	Type	O	C	K	W	Avg
CSFU [8]	OCDA	36.2	34.9	32.4	27.6	32.8
DACS [25]	UDA	36.8	37.0	37.4	28.8	35.0
RobustNet [4]	DG	37.1	38.3	40.1	29.6	36.3
DHC [19]	OCDA	38.9	38.0	40.6	30.0	36.9
Ours (w/o Updating)	OCDA	41.5	40.3	42.7	30.1	38.7
Ours (w/ Updating)	OCDA	42.6	41.1	43.4	30.9	39.5

multi-discriminator to minimize the domain gaps. In comparison, the latest UDA approaches DACS and SAC show 36.6% and 36.4%, outperforming both CDAS and CSFU. The reason behind is that both DACS and SAC adopt various self-supervision techniques to minimize the domain gaps, which proves to be more effective than GAN-based approaches. In comparison, the proposed approach demonstrates effectiveness on this benchmark dataset with 40.2% of mean IoU over all classes.

We present experimental results on SYNTHIA→C-Driving shown in Table 1b. We consider the 11 classes for final evaluation. The proposed method achieves 40.0% of mean IoU over the 11 classes. For other OCDA approaches, DHA achieves 37.6%, CSFU produces 34.8%, and CDAS generates 34.0% of mean IoU. Moreover, the UDA approaches DACS and SAC generate 36.5% and 36.3% of mean IoU. Our approach outperforms all the existing OCDA approaches and the latest UDA approaches.

Generalization to the Open Domains. We also evaluate the domain generalization of the proposed approach against existing UDA and OCDA approaches. The results are presented in Table 2a and 2b. Our work is compared with the latest domain generalization (DG) approach RobustNet [4]. For all the UDA and OCDA approaches, we first train them with the labeled source and the unlabeled target images, and we evaluate their performance with the validation of the open domains. RobustNet generates 37.0% of mean IoU in Table 2a and 36.3% of mean IoU in Table 2b. Note that RobustNet only requires labeled source data during training. This shows that the DG approach is more effective in generalizing to the open domains than the existing UDA and OCDA approaches DACS and CSFU. Without any online updating, our approach achieves 40.0% of mean IoU in Table 2a and 38.7% of mean IoU in Table 2b. Our approach outperforms all the UDA approaches, OCDA approaches, and the DG approach listed in the table. The reason might be that our approach is more powerful for learning the domain invariant features which improve the generalization of the model toward novel

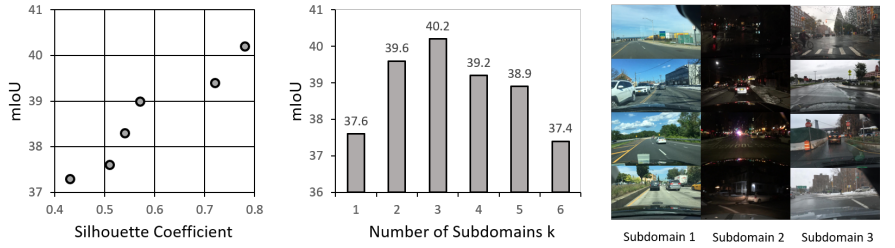

Fig. 3: We conduct the ablation study on the proposed automatic domain separation using GTA5→C-Driving with ResNet101 backbone. (a) The scatterplot shows the correlation between our approach's mean IoU and the Silhouette Coefficient score. (b) The mean IoU of our approach with different number of subdomains k. (c) The sample images from the subdomains of the C-Driving dataset.

domains. The performance gain of our approach with updating further shows the efficacy of the proposed online updating with consistency regularization.

5.3 Ablation Study

Generating Optimal Subdomains. We first conduct the ablation study on the correlation between the mean IoU of the proposed approach with Silhouette Coefficient (SC) score on the subdomain separation in Figure 3(a). It shows a positive correlation, which means that the SC score is effectively finds the optimal number of subdomains for the compound target domain. Moreover, we evaluate the mean IoU score with the different number of subdomains k in Figure 3(b). Finally, we set up $k = 3$ and present the sample images from the subdomains of the C-Driving dataset in Figure 3(c). We also evaluate the efficacy of the proposed subdomain style purification (SSP) in Table 3b. Without using SSP, the performance drops 0.5% of mean IoU.

Multi-teacher and Single Model. The ablation study on the multi-teacher learning of our proposed approach is presented in Table 3a and Table 3b. Applying a single model in our approach delivers 38.0% of mean IoU, leading to the the most significant drop 2.2%, shown in Table 3b. We further combine DACS with multi-teacher learning, and the mean IoU reaches from 36.6% to 39.1%. We argue that utilizing a single model is less effective than the multi-teacher models. Because adapting to one subdomain might hinder the single model from adapting to other dissimilar subdomains. Thus, we employ a multi-teacher framework in which each teacher adapts to one subdomain separately. And the multiple teachers together provide a comprehensive guide to the student model to adapt to all the target subdomains. We further present the qualitative results about

Table 3: The ablation study on the efficacy of the components of our model. (a) We compare with one baseline model DACS [25] and evaluate the performance gain of the bidirectional photometric mixing and the multi-teacher learning. (b) We evaluate the performance drop of our model by removing each component from it. Our model is trained GTA5→C-Driving with ResNet101 backbone and tested on C-Driving validation set.

(a) The performance gain.

GTA5→C-Driving	
Model	mIoU
DACS [25]	36.6
DACS + Multi-teacher Learning	39.1
DACS + Bidirectional Mixing	37.3
DACS + Photometric Mixing (Γ, Δ)	37.4
DACS + Bidirectional Photometric Mixing	37.8
Ours	40.2

(b) The performance drop.

GTA5→C-Driving		
Configuration	mIoU	Gap
w/o Multi-teacher Learning	38.0	-2.2
w/o Mixing on One Direction ($\alpha = 0$)	38.5	-1.7
w/o Mixing on One Direction ($\beta = 0$)	38.9	-1.3
w/o Subdomain Style Purification	39.7	-0.5
w/o Adaptive Distillation	39.6	-0.6
Full Framework	40.2	-

Fig. 4: We compare the mixed images from the source domain to the target domain. (a) the source image; (b) the target image; (c) the mixed images without using photometric transform, and the style inconsistency exists; (d) the mixed images using photometric transform, and the style inconsistency is mitigated; (e) the mask to crop the source image.

the target image prediction maps from each subdomain by the multi-teachers and the single-teacher model in Figure 5.

Bidirectional Photometric Mixing. We further conduct the ablation study for the bidirectional photometric mixing (BPM), shown in Table 3a and Table 3b. Our model is trained on GTA5 → C-Driving with ResNet101 backbone and tested on C-Driving validation set. By making $\alpha = 0$ to remove the mixing on one direction (ClassMix), the mean IoU drops 1.7%, while making $\beta = 0$ to remove the other directional mixing (CutMix), it decreases by 1.3%. This suggests that ClassMix contributes slightly more to the final performance. We also use the baseline model DACS for an in-depth analysis. We add the bidirectional photometric mixing with the DACS, the performance increase from 36.6% to

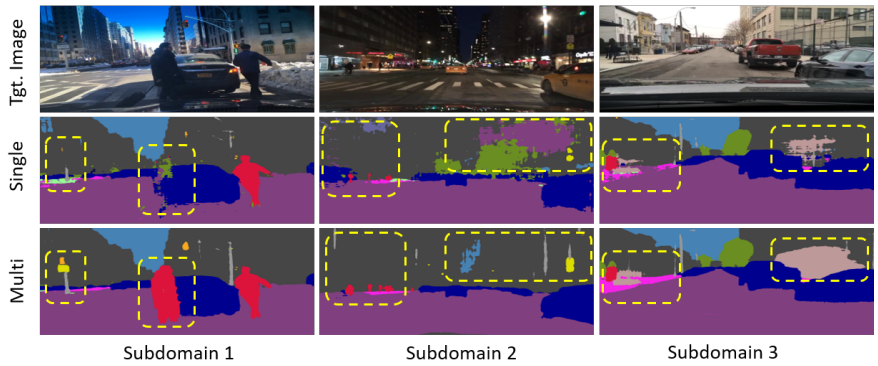

Fig. 5: We present the predicted segmentation maps of the target images from every target subdomain. The maps in the second row are generated using a single model. The maps in the third row are generated using the multi-teacher models.

37.8% shown in Table 3a; we then combine DACS with only bidirectional mixing, the mean IoU rise up to 37.3%; we further add DACS with only photometric transform on mixing (use Γ and Δ), the mean IoU reaches to 37.4%. The reason behind is that DACS utilizes a simple mixing method that contains only one direction and generates the mixed image with the style inconsistency inside. However, we propose a bidirectional mixing scheme and apply the photometric transform to mitigate the style inconsistency on the generated images. We present the qualitative results to show this issue in Figure 4. The style inconsistency is mitigated in Figure 4(d) compared with Figure 4(c) on the mixing direction from the source domain to the target domain.

6 Conclusion

Open compound domain adaptation (OCDA) considers the target domain as the compound of multiple unknown subdomains. In this work, we first propose automatic domain separation to find the optimal number of subdomains. Then we design a multi-teacher framework with bidirectional photometric mixing to align the domain gap between the source domain and the compound target domain, and we further evaluate its generalization to novel domains. Our current work is only focused on segmentation task and we leave the study on other visual tasks for future research.

7 Acknowledgment

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20224000000100).

References

1. Abu Alhaija, H., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets computer vision: Efficient data generation for urban driving scenes. IJCV **126**(9), 961–972 (2018)
2. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. In: CVPR. pp. 15384–15394 (2021)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4), 834–848 (2017)
4. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: CVPR. pp. 11580–11590 (2021)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009)
7. French, G., Oliver, A., Salimans, T.: Milking cowmask for semi-supervised image classification. arXiv preprint arXiv:2003.12022 (2020)
8. Gong, R., Chen, Y., Paudel, D.P., Li, Y., Chhatkuli, A., Li, W., Dai, D., Van Gool, L.: Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In: CVPR. pp. 8344–8354 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
10. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML. pp. 1989–1998. PMLR (2018)
11. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In: CVPR. pp. 6891–6902 (2021)
12. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: CVPR. pp. 603–612 (2019)
13. Liu, Z., Miao, Z., Pan, X., Zhan, X., Lin, D., Yu, S.X., Gong, B.: Open compound domain adaptation. In: CVPR. pp. 12406–12415 (2020)
14. Ma, H., Lin, X., Wu, Z., Yu, Y.: Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In: CVPR. pp. 4051–4060 (2021)
15. Olsson, V., Tranheden, W., Pinto, J., Svensson, L.: Classmix: Segmentation-based data augmentation for semi-supervised learning. In: WACV. pp. 1369–1378 (2021)
16. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: ECCV. pp. 762–780. Springer (2020)
17. Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S.: Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: CVPR. pp. 3764–3773 (2020)
18. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV. pp. 464–479 (2018)
19. Park, K., Woo, S., Shin, I., Kweon, I.S.: Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. In: NeurIPS (2020)

20. Rafael, C.G., Richard, E.W., Steven, L.E., Woods, R., Eddins, S.: Digital image processing using MATLAB. Tata McGraw-Hill (2010)
21. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV. pp. 102–118. Springer (2016)
22. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016)
23. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. JCAM **20**, 53–65 (1987)
24. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV. pp. 10765–10775 (2021)
25. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: Dacs: Domain adaptation via cross-domain mixed sampling. In: WACV. pp. 1379–1389 (2021)
26. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
27. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR. pp. 7472–7481 (2018)
28. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019)
29. Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W.m., Huang, T.S., Shi, H.: Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In: CVPR. pp. 12635–12644 (2020)
30. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: ICCV. pp. 2100–2110 (2019)
31. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: CVPR. pp. 6023–6032 (2019)
32. Zendel, O., Honauer, K., Murschitz, M., Steininger, D., Dominguez, G.F.: Wilddash-creating hazard-aware benchmarks. In: ECCV. pp. 402–416 (2018)
33. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. CVPR (2021)
34. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR. pp. 6848–6856 (2018)
35. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
36. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
37. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018)