This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Open-Set Domain Adaptation for Semantic Segmentation

Seun-An Choe* Ah-Hyung Shin* Keon-Hee Park Jinwoo Choi† Gyeong-Moon Park†

Kyung Hee University, Yongin, Republic of Korea

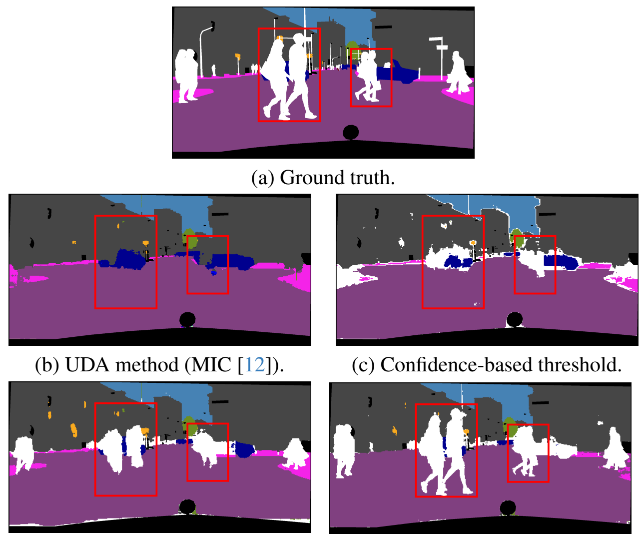
{dragoon0905, dkgud111, pgh2874, jinwoochoi, gmpark}@khu.ac.kr

Abstract

*Unsupervised domain adaptation (UDA) for semantic segmentation aims to transfer the pixel-wise knowledge from the labeled source domain to the unlabeled target domain. However, current UDA methods typically assume a shared label space between source and target, limiting their applicability in real-world scenarios where novel categories may emerge in the target domain. In this paper, we introduce Open-Set Domain Adaptation for Semantic Segmentation (OSDA-SS) for the first time, where the target domain includes unknown classes. We identify two major problems in the OSDA-SS scenario as follows: 1) the existing UDA methods struggle to predict the exact boundary of the unknown classes, and 2) they fail to accurately predict the shape of the unknown classes. To address these issues, we propose **B**oundary and **U**nknown **S**hape-Aware open-set domain adaptation, coined **BUS**. Our BUS can accurately discern the boundaries between known and unknown classes in a contrastive manner using a novel dilation-erosion-based contrastive loss. In addition, we propose OpenReMix, a new domain mixing augmentation method that guides our model to effectively learn domain and size-invariant features for improving the shape detection of the known and unknown classes. Through extensive experiments, we demonstrate that our proposed BUS effectively detects unknown classes in the challenging OSDA-SS scenario compared to the previous methods by a large margin. The code is available at `https://github.com/KHU-AGI/BUS`.*

1. Introduction

In semantic segmentation, a model predicts pixel-wise category labels given an input image. Semantic segmentation has a lot of applications, e.g., autonomous driving [1], human-machine interaction [2], and augmented reality. Over the past decade, there has been notable advancement in supervised semantic segmentation driven by deep neural networks [3–6]. However, supervised semantic segmentation requires pixel-level annotations, which are labor-intensive and costly to collect. To mitigate the challenges, unsupervised domain adaptation (UDA) has emerged. Many studies [7–12] leverage the already-labeled source data to achieve high performance on the unlabeled target data. Notably, synthetic datasets such as GTA5 [13] and SYNTHIA [14] which are automatically generated by game engines present valuable resources for UDA research.



(a) Ground truth.
(b) UDA method (MIC [12]).
(c) Confidence-based threshold.
(d) Unknown head-expansion.
(e) BUS (**Ours**).

Figure 1. Visualization of prediction maps in the OSDA-SS scenario. The pixels detected by the white color mean the unknown classes. The naive UDA method (b) is completely unaware of the *unknown* classes. Even after applying simple techniques to help the UDA model recognize the *unknown*, it still struggles to accurately predict the shape of the *unknown*, as shown in (c) and (d).

UDA methods typically presume that source and target domains share the same label space. Such an assumption is not reasonable in real-world applications. In the target data, novel categories not presented in the source dataset (target-private categories) may emerge, leading to an Open-

*Equal contribution
†Corresponding authors

Set Domain Adaptation (OSDA) setting. The conventional UDA method may significantly fail under the OSDA setting, e.g., a model erroneously label a person walking on the road as the road itself as shown in Figure 1(b). The desired model should reject any target-private classes as *unknown* rather than misclassifying it as a known class. While OSDA has been widely explored in image classification [15–18], its application to semantic segmentation remains unexplored to the best of our knowledge. In this work, we tackle the interesting and challenging problem of Open-Set Domain Adaptation for Semantic Segmentation (OSDA-SS). Here, we deal with the labeled source data and the unlabeled target data containing classes not found in the source domain. In the OSDA-SS setting, the goal is to accurately predict pixel-wise category labels in the target domain and correctly distinguish the classes not seen during training as *unknown*.

One can design reasonable baselines by extending well-established UDA methods. One approach could be a confidence-threshold baseline. We train a model by using the UDA algorithm without considering target-private classes. During inference, the model identifies pixels with confidence scores below a predefined threshold as *unknown*. We show the predicted segmentation map from the confidence-threshold baseline in Figure 1(c). Another baseline could be a head-expansion baseline. We expand the classification head from C to $(C + 1)$ dimensions, where C represents the number of known classes. During training, when generating pseudo labels, we assign pixels with confidence scores lower than a specific threshold to the $(C+1)$-th head and train with the pseudo labels. We show the predicted segmentation map from the head-expansion baseline in Figure 1(d). These baselines sometimes reject target-private classes as *unknown*, but they often fail to do so, resulting in poor performance on the target dataset.

In this work, we build a model upon the head-expansion baseline. We find two failure modes of the baseline and propose a novel **B**oundary and **U**nknown **S**hape-Aware (**BUS**) OSDA-SS method. First, the previous models are often less confident or even fail near the boundaries of objects [19–21]. We find that the problem is even more severe for target-private classes due to lack of supervision. To address this issue, we propose a new **D**ilation-**E**rosion-based **CON**trastive (**DECON**) loss that manifests the boundaries through morphological operations, specifically dilation and erosion. Given a target image, we generate a target private mask using pseudo-labeling with the expanded head. Subsequently, we generate a boundary mask by subtracting the original private mask from the dilated private mask, indicating the region of *known* classes near the boundaries. We generate an erosion mask by applying erosion to the private mask, indicating more confident regions of the *private classes*. We then train the model in a contrastive manner using the features from the erosion mask and the boundary mask as positive and negative samples, respectively. With DECON loss, our model clearly discerns the common and private classes near the boundaries.

Second, the baseline model faces challenges in accurately predicting the shape of *unknown*. If the model consistently predicts the same object regardless of variations in size, it indicates that the model relies more on shape information than size information to recognize the object. Inspired by this motivation, we propose a new data mixing augmentation, **OpenReMix**. This method involves 1) resizing a random thing class from the source image and mixing it with the target image during training to consistently predict the same object even when its size varies. In addition, since there are no *unknown* classes in the source image, 2) we cut the parts predicted as *unknown* from a target image and paste them into a source image for supplemental learning of the last $(C+1)$-th head, aiding in the rejection of *unknown* during source training. This delicate mixing strategy notably enhances the detection capability of *unknown*, with a specific emphasis on capturing the shape information. By addressing the failure modes, the proposed BUS achieves significant performance gains on public benchmarks: GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes.

We summarize our major contributions as follows:

- To the best of our knowledge, we introduce a new task, Open-Set Domain Adaptation for Semantic Segmentation (OSDA-SS) for the first time. To tackle this challenging task, we propose a novel **B**oundary and **U**nknown **S**hape-Aware OSDA-SS method, coined **BUS**.
- We introduce DECON loss, a new dilation-erosion-based contrastive loss to address the less confident and wrong predictions near the class boundaries.
- We propose OpenReMix, which leads our model to learn size-invariant features and leverages *unknown* objects from target to source to train the expanded head efficiently. OpenReMix encourages our model to focus on shape information of *unknown* classes.
- We conduct extensive experiments to validate the effectiveness of our proposed method. The proposed BUS shows state-of-the-art performance on public benchmark datasets with a significant margin.

2. Related Work

2.1. Semantic Segmentation.

Semantic segmentation, which is a task to predict pixel-wise labels from the input images, has witnessed significant advances over the last decade. Key developments include fully convolution networks (FCNs) [3], dilated convolution [4, 5], global pooling [22], pyramid pooling [23–25], and attention mechanism [26–29]. Despite their success, these methods typically depend on a large amount of labeled data which is label-intensive and costly to collect. In contrast, we formulate the semantic segmentation problem as domain adaptation to mitigate the annotation cost.

2.2. Unsupervised Domain Adaptation for Semantic Segmentation.

Recently, there has been a lot of work on unsupervised domain adaptation (UDA) for semantic segmentation. UDA methods for semantic segmentation generally fall into two categories: adversarial learning-based and self-training approaches. Adversarial learning-based methods [7, 30–35] utilize an adversarial domain classifier to learn domain-invariant representations, aiming to deceive the domain classifier. Self-training methods [9–12, 36–44] create pseudo labels for each pixel in the target domain image using confidence thresholding. Several self-training methods iteratively re-train the models, which result in enhanced performance on the target domain. Despite the great success, most previous works assume a closed set setting, where the source and target domains share the same label space. In this work, we relax this unrealistic assumption and tackle the problem of open-set domain adaptation for semantic segmentation (OSDA-SS). To the best of our knowledge, there is no prior work to tackle this problem.

2.3. Open-Set Domain Adaptation

Open-set domain adaptation (OSDA) extends UDA to handle novel categories in the target domain that are not present in the source domain. The primary goal of OSDA is to effectively distinguish the unknown categories from the known classes while reducing the domain gap between the source and target domains. Several OSDA methods have been proposed for the classification task [15, 17, 45–47]. However, in semantic segmentation task, which requires a higher degree of spatial information compared to classification, directly applying classification methods struggles to effectively differentiate unknown categories. The most similar work [48] to our method also deals with the novel classes that do not exist in the source domain. However, it accesses pre-defined private category definitions. To address this challenge, we propose a novel OSDA-SS task to discriminate unknown categories without needing to know any information about pre-defined class definitions.

2.4. Domain Mixing Augmentation.

To improve the generalization power of deep neural networks, mixup [49, 50] and its variants [44, 51–59] have been proposed. Especially, domain mixing augmentation demonstrates significant performance improvement in UDA [44, 51–54, 60] by utilizing domain-mixed images as training data to encourage learning of domain-invariant feature representations. We propose OpenReMix, aiming to empower our model in capturing shape information, notably for the *unknown* classes.

3. Method

3.1. Problem Formulation

In this section, we formulate a novel OSDA-SS task for the first time. In OSDA-SS, a network is trained with the source images $X_s = \{x_s^1, x_s^2, ..., x_s^{i_s}\}$ and the corresponding labels $Y_s = \{y_s^1, y_s^2, ..., y_s^{i_s}\}$ to ensure effective performance in the target domain $X_t = \{x_t^1, x_t^2, ..., x_t^{i_t}\}$ without labels. $x_s^{i_s} \in \mathbb{R}^{3\times H\times W}$ and $y_s^{i_s} \in \mathbb{R}^{C\times H\times W}$ are the i_s-th source domain image and the pixel-wise label. H and W are the height and width of the image, respectively, and C denotes the number of categories in the source domain. In the target domain, we only have the image $x_t^{i_t} \in \mathbb{R}^{3\times H\times W}$ without the corresponding labels. The source and target domains share C categories, and the target domain has additional unknown classes, *i.e.*, the target images contain unknown objects. In this setting, the goal of OSDA-SS is to train a segmentation model f_θ using both the labeled source data (X_s, Y_s) and the unlabeled target data X_t, and eventually the learned model f_θ should predict both known and unknown classes well on the target domain.

3.2. Baseline

Inspired by the UDA methods based on self-training [10–12, 44], we build a OSDA-SS baseline by extending the number of classifier heads from C to $(C+1)$, where the $(C+1)$-th head corresponds to *unknown* classes. The segmentation network f_θ is trained with the labeled source data using the following categorical cross-entropy loss $\mathcal{L}^s_{seg}$:

$$\mathcal{L}^s_{seg} = -\sum_{j=1}^{H\cdot W}\sum_{c=1}^{C+1} y_s^{(j,c)} \log f_\theta(x_s)^{(j,c)}, \quad (1)$$

where $j \in \{1, 2, ..., H \cdot W\}$ denotes the pixel index and $c \in \{1, 2, ..., C+1\}$ denotes the class index. To alleviate the domain gap between the source and the target domains, the baseline utilizes a teacher network g_ϕ to generate the target pseudo-labels. The pseudo-label $\hat{y}_{tp}^{(j)}$ for the j-th pixel considering *unknown* is acquired as follows:

$$\hat{y}_{tp}^{(j)} = \begin{cases} c', & \text{if } \left(\max_{c'} g_\phi(x_t)^{(j,c')} \geq \tau_p\right) \\ C+1, & \text{otherwise} \end{cases}, \quad (2)$$

where $c^{'} \in \{1, 2, ..., C\}$ denotes a class belonging to known classes and τ_p is a threshold. Using the above equation, we assign the less confident pixels as the *unknown* class when the maximum softmax probability is lower than τ_p. Since we cannot completely trust the pseudo-labels above, we estimate the confidence of the pseudo-label by utilizing the

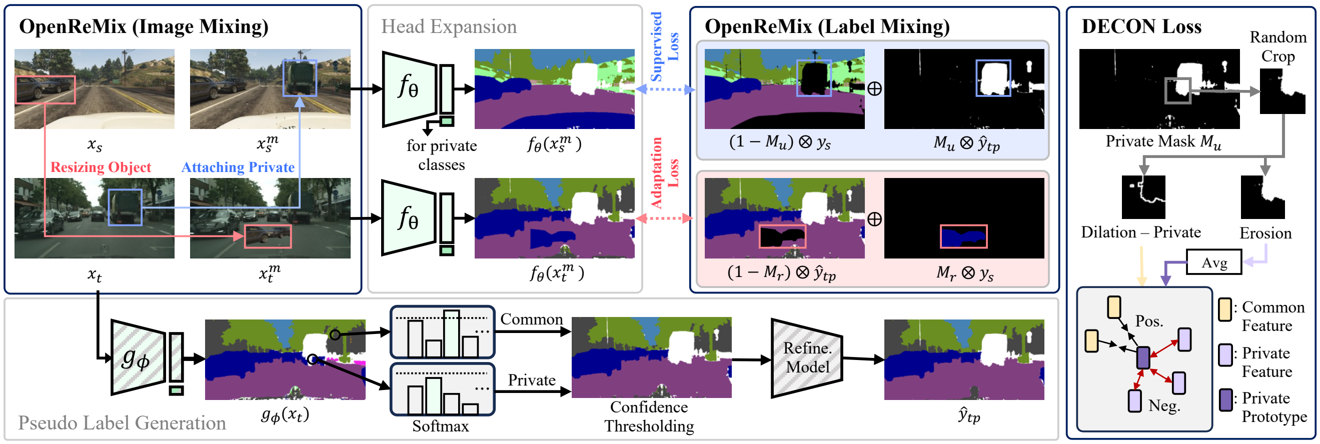


Figure 2. Overview of our proposed Boundary and Unknown Shape-Aware (BUS) method. We generate the mixed source image x_s^m and the mixed target image x_t^m from OpenReMix. The model is trained using the mixed source label and the mixed target pseudo-labels with supervised loss and adaptation loss, respectively. Especially, the expanded head is trained with the parts that predicted as unknown in pseudo-labels. Pseudo-labels are generated by thresholding the softmax probability and passing through the refinement network. DECON loss utilizes the dilation and erosion operations to distinguish the known and unknown classes near the boundaries.

ratio of confident pixels [44]. To this end, we count the number of pixels that have the maximum probability values exceeding a certain threshold τ_t as follows:

$$q_t = \frac{1}{H \cdot W} \sum_{j=1}^{H \cdot W} \left[\max_{c'} g_\phi(x_t)^{(j,c')} \geq \tau_t \right], \tag{3}$$

where q_t means the confidence of the pseudo-label for the image. The network f_θ is trained using the pseudo-labels and the corresponding confidence estimates with the following categorical cross-entropy loss $\mathcal{L}_{seg}^t$:

$$\mathcal{L}_{seg}^t = - \sum_{j=1}^{H \cdot W} \sum_{c=1}^{C+1} q_t \hat{y}_{tp}^{(j,c)} \log f_\theta(x_t)^{(j,c)}. \tag{4}$$

Finally, we update the teacher network g_ϕ from f_θ using the exponential moving average (EMA) [63] with a smoothing factor α at the $(t+1)$-th iteration, where the equation is shown as follows:

$$\phi_{t+1} = \alpha \phi_t + (1-\alpha)\theta_t. \tag{5}$$

Based on this baseline, we propose a novel **B**oundary and **U**nknown **S**hape-Aware OSDA method, coined **BUS**, which involves a new loss function to manifest the boundaries of known and unknown classes (see Section 3.3) and a new domain mixing augmentation to detect the shape of unknown objects robustly (see Section 3.4).

3.3. Dilation-Erosion-based Contrastive Loss

Semantic segmentation models often struggle to confidently predict object boundaries [19–21], especially for target-private classes, where the absence of label information makes boundary prediction even more challenging. Since the models predict the boundaries with low confidence estimates, the quality of the generated pseudo-labels may not be accurate. If the model can confidently identify the boundaries of unknown classes, accurate predictions of unknown classes become feasible.

To discern the boundaries effectively, we leverage two morphological operations, which are dilation and erosion. First, we utilize the pseudo-labels of the target image to create a target private mask as follows:

$$M_u^{(j)} = \begin{cases} 1, & \text{if } \hat{y}_{tp}^{(j)} = C+1 \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where j denotes the pixel index. Next, we apply the dilation function $h_d(\cdot)$ and the erosion function $h_e(\cdot)$ to the randomly cropped target private mask, generating dilation and erosion masks. In the dilation mask, we subtract the original target private mask to identify the regions associated with the common classes near the boundaries. On the other hand, the erosion mask emphasizes the regions that definitively belong to the private class. We generate these masks by the following equations:

$$M_N = h_d(M_u') - M_u', \tag{7}$$

$$M_P = h_e(M_u'), \tag{8}$$

where $M_u' = r(M_u)$ and $r(\cdot)$ is a function of random crop. M_N and M_P denote the masks representing the common and private parts, respectively. To construct a contrastive loss, we generate anchor, positive, and negative samples using these masks as follows:

$$z_i = \text{avg}(M_P \odot f_\theta(x_t)), \tag{9}$$

$$z_j = M_P \odot f_\theta(x_t), \tag{10}$$

$$z_k = M_N \odot f_\theta(x_t), \tag{11}$$

where z_i is an anchor, $\text{avg}(\cdot)$ denotes the average pooling layer, and z_j and z_k represent positive and negative samples, respectively. We utilize z_i as a prototype calculated by the average of positive samples. Finally, we define the contrastive loss [64] using z_i, z_j, and z_k as follows:

$$\mathcal{L}_{DECON} = -\log\left[\sum_{p=1}^{N_p}\exp(z_i \cdot z_j^p/\tau)/\sum_{n=1}^{N_n}\exp(z_i \cdot z_k^n/\tau)\right], \tag{12}$$

where τ is a temperature parameter. N_n and N_p denote the number of negative and positive pixels. To sum up, the proposed $\mathcal{L}_{DECON}$ allows our model to better distinguish between common and private classes near the boundaries.

3.4. OpenReMix

Resizing Object. We identify that the head-expansion baseline model fails to accurately predict the shape of the private classes. We hypothesize that if a model consistently predicts the same object regardless of size variations, the model can accurately predict the shape of the object as well. To this end, we extend the domain mixing method Classmix [51], which selects half of the classes from the source and appends them to the target image to learn domain-invariant features. On top of the Classmix, we introduce an additional step where we select one more thing class from the source image, resize it, and paste it to the random location of a target image with resizing object mask M_r. The mixed target image contains the same objects as the source image, but the sizes of the objects are different. Therefore, the model learns not only domain-invariant representations but also size-invariant representations from the mixed target images and the source images. This extension enhances the robustness of the model to size variations, contributing to the accurate prediction of the shape of unknown classes leading to superior open-set domain adaptation performance.

Attaching Private. As described in Section 3.2, to address the target private classes, we expand the segmentation head. The expanded head is trained with the target pseudo-labels which contain the private labels. However, since there are no private classes in the source image, we cannot utilize the source data to update the additional head of the model. To overcome this inefficiency in training, we copy the parts of target private classes and paste them into a source image. Given a target image, we create a target private mask M_u as Eq. (6). With the target private mask, we copy the private regions in the target image to a source image, resulting in a private class-mixed source image. Similarly, by combining the labels of the source and the pseudo-labels of the target, we generate mixed source labels. This augmentation offers a significantly larger dataset for training to reject private classes, leading to improved open-set domain adaptation performance. We formalize the attaching private process as follows. We generate a mixed source image x_s^m and the corresponding source label y_s^m using the following equations:

$$x_s^m = M_u \odot x_t + (1 - M_u) \odot x_s, \tag{13}$$
$$y_s^m = M_u \odot \hat{y}_{tp} + (1 - M_u) \odot y_s, \tag{14}$$

where x_t and $\hat{y}_{tp}$ denote the target image and its pseudo-label. The mixed image x_s^m and the mixed label y_s^m are applied to Eq. (1), instead of the source image x_s and the corresponding label y_s.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluated our framework over two challenging synthetic-to-real scenarios in autonomous driving, i.e., GTA5 → Cityscapes and SYNTHIA → Cityscapes. GTA5 [13] is a synthesized dataset, which consists of 24,966 images with a resolution of 1914×1052. SYNTHIA [14] is also a synthesized dataset, which contains 9,400 images with resolution 1280×760. Cityscapes [65] is a real-image dataset with 2,975 training samples and 500 validation samples with resolution 2048×1024. It shares 19 classes with GTA and 16 classes with SYNTHIA.

Scenario Construction. Using these datasets, we established new scenarios tailored for the OSDA-SS task. First, to create new classes emerging in the target domain that are not present in the source domain, we selected certain source classes to be removed. In autonomous driving scenarios, the classes that are likely to emerge in the target domain are expected to be “thing” classes. Stuff classes representing the background area typically do not emerge as new classes. Therefore, we selected specific classes from the “thing” categories to be excluded. The following list denotes the classes designated as unknown in GTA5 and SYNTHIA.

- GTA5: “pole”, “traffic sign”, “person”, “rider”, “truck”, and “train”.
- SYNTHIA: “pole”, “traffic sign”, “person”, “rider”, “truck”, “train”, and “terrain”.

Notably, SYNTHIA includes the “terrain”, which inherently lacks labels from the outset. Second, in order to avoid training the excluded classes, pixels corresponding to those classes were designated as “ignore” and were not included in the loss function during training. Finally, during the evaluation of the target domain, the above classes were treated as single unknown class.

Evaluation Metrics. Previous works [9–12] used mIoU (mean Intersection-over-Union) as the evaluation metric,

Method	Road	S.walk	Build.	Wall	Fence	Light	Veget.	Terrain	Sky	Car	Bus	M.bike	Bike	Common	Private	H-Score
GTA5 → Cityscapes																
OSBP [15]	4.92	3.93	42.8	2.55	6.04	14.29	68.58	26.50	44.21	41.78	0.94	7.20	3.42	20.55	4.49	7.34
UAN [18]	65.97	23.41	76.41	37.26	18.50	20.13	80.57	30.37	82.47	77.35	27.80	16.62	0.00	38.00	3.59	6.56
UniOT [17]	17.67	5.14	44.86	**55.45**	2.31	52.61	40.01	3.37	79.43	52.87	52.31	7.18	0.00	20.20	5.36	7.49
ASN [7]	82.34	2.21	75.30	8.01	3.52	9.99	71.96	15.61	70.97	77.16	22.59	20.8	0.06	35.43	10.84	16.60
Pixmatch [9]	79.27	2.06	72.36	6.96	2.94	11.07	76.29	23.23	77.72	79.77	44.72	18.02	0.01	38.03	9.46	15.15
DAF [10]	94.26	48.69	83.47	38.67	32.83	41.71	87.79	39.15	**93.59**	85.29	47.04	28.36	46.86	61.26	14.63	23.36
HRDA [11]	**95.14**	62.58	82.92	47.44	43.57	53.18	88.26	44.42	92.92	90.23	57.43	14.71	56.83	63.82	12.13	20.39
MIC [12]	93.26	58.96	79.30	21.62	31.41	39.32	85.48	31.94	91.64	88.16	44.77	47.64	42.77	58.17	11.87	19.71
BUS **(Ours)**	95.06	**66.65**	**90.53**	55.37	**55.38**	**57.20**	**91.12**	**49.69**	92.96	**93.50**	**68.81**	**58.73**	**67.04**	**72.47**	**55.42**	**62.81**

Method	Road	S.walk	Build.	Wall	Fence	Light	Veget.	Sky	Car	Bus	M.bike	Bike	Common	Private	H-Score
SYNTHIA → Cityscapes															
OSBP [15]	6.71	9.49	49.83	0.70	0.0	0.76	26.03	36.91	20.04	4.76	2.90	8.70	13.20	4.90	7.14
UAN [18]	33.24	19.03	71.49	4.02	0.05	14.34	75.78	81.06	53.88	19.34	8.14	21.84	31.30	4.53	7.91
UniOT [17]	0.00	16.79	18.52	1.05	6.49	16.8	14.52	57.4	6.48	2.59	3.73	3.88	12.35	5.49	7.06
ASN [7]	72.70	41.29	73.59	7.38	0.08	1.17	71.35	82.22	67.35	23.30	0.94	20.56	38.49	4.62	8.25
Pixmatch [9]	74.16	8.15	76.21	0.01	0.0	5.64	44.15	63.76	44.66	17.27	0.13	0.38	26.30	6.87	11.00
DAF [10]	70.10	39.65	83.09	22.75	4.66	41.19	81.56	91.79	84.36	51.13	43.78	46.20	51.49	9.07	15.57
HRDA [11]	85.62	41.74	83.29	36.35	0.86	35.17	83.98	90.90	84.74	50.42	46.78	58.33	54.68	12.68	20.82
MIC [12]	**88.31**	**70.71**	85.00	26.23	**6.60**	35.27	84.80	91.41	81.47	53.62	55.39	58.20	57.46	10.02	17.23
BUS **(Ours)**	86.85	43.49	**89.35**	**46.12**	4.39	**54.29**	**87.90**	**92.49**	**91.46**	**61.23**	**58.11**	**59.81**	**64.62**	**33.37**	**44.01**

Table 1. Performance on two different benchmarks. Our proposed BUS achieved the state-of-the-art performance with remarkable improvement in H-Score +39.45% against DAFormer in GTA → Cityscapes and +23.19% against HRDA in SYNTHIA → Cityscapes.

which averaged the IoU of each class. Since we treated every unknown classes as single unknown class, simply averaging would diminish the impact of private classes significantly. Therefore, inspired by [66], we utilized the harmonic mean of the mean IoU score for known classes (common) and the IoU score for one unknown class (private) as our evaluation metric, known as the H-Score.

Implementation Details. We adopted DAFormer [10] network with the MiT-B5 encoder [6] pre-trained on imageNet-1K [67]. We followed the multi-resolution self-training strategy and training parameters of MIC [12]. The network was trained with AdamW [68]. The learning rates were set to 6e-5 for the backbone and 6e-4 for the decoder head, with a weight decay of 0.01 and linear learning rate warm-up over 1.5k steps. EMA factor was α=0.999. We utilized the Rare Class Sampling [10], ImageNet Feature Distance [10], DACS [44] data augmentation, and Masked Image Consistency module [12]. We trained on a batch of two 512 $\times$ 512 random crops for 40k iterations. We used MobileSAM [69] for the refinement model. The refinement process is described in the supplemental material.

Baselines. We compared our approach with two scenarios. The first scenario comprised the Open-Set Domain Adaptation (OSDA) method like OSBP [15] and Universal Domain Adaptation (UniDA) methods like UAN [18] and UniOT [17], which were capable of rejecting unknown classes but were primarily designed for classification tasks. The second scenario was Unsupervised Domain Adaptation (UDA) methods for semantic segmentation in closed-set setting, which included AdaptSegNet (ASN) [7], Pixmatch [9], DAFormer (DAF) [10], HRDA [11], and MIC [12]. In the UDA method, we assigned the unknown label for regions with low confidence scores during inference. For OSDA and UniDA methods, we replaced the classification network with the DeepLabv2 [5] segmentation network, which uses ResNet-101 [70] as the backbone, and adopted the image-level methods to the pixels.

4.2. Comparison with the State-of-the-Art

Table 1 showed the experimental results of GTA5 → Cityscapes and SYNTHIA→ Cityscapes, respectively. The classification methods struggled to accurately discriminate the private classes in semantic segmentation tasks, which demanded a higher degree of spatial information. The UDA methods also faced challenges in effectively distinguishing private classes when simply leveraging a confidence-based approach. In contrast, our proposed approach significantly outperformed the other comparison methods in H-Score. Especially, compared to the best baseline, our proposed BUS achieved a performance improvement of about

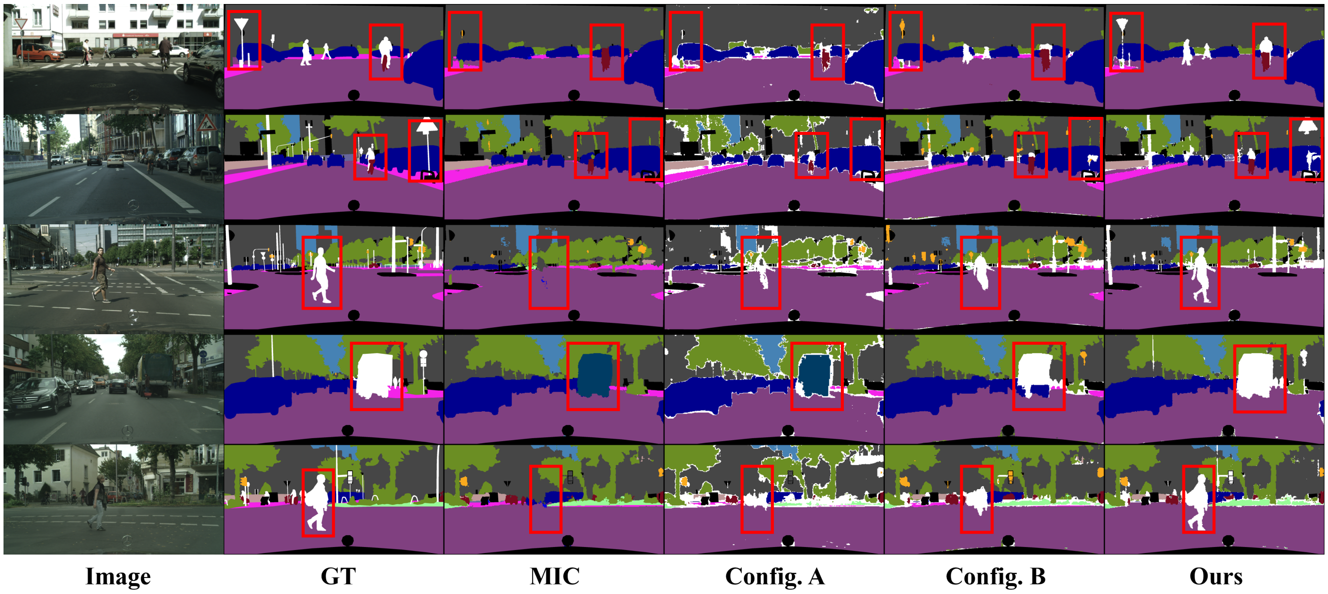


Figure 3. Qualitative comparison of our method with MIC, confidence-based MIC (Config. A), and head-expansion (Config. B) on the GTA5 $\rightarrow$ Cityscapes. GT represents the ground truth.

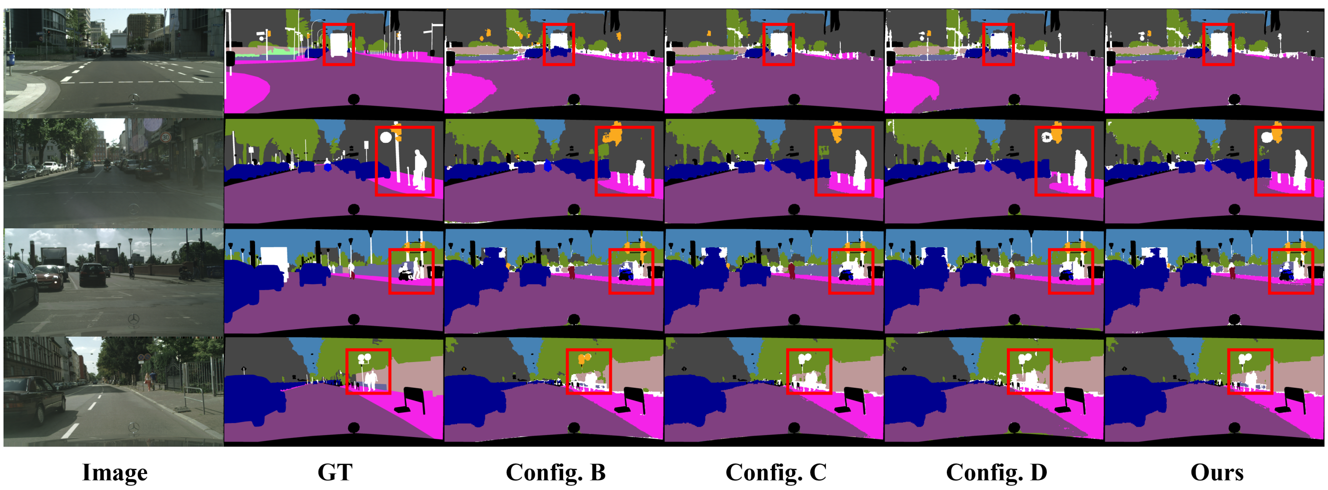


Figure 4. Qualitative comparison of our method with head-expansion (Config. B), DECON loss (Config. C), and OpenReMix (Config. D) on the SYNTHIA $\rightarrow$ Cityscapes. GT represents the ground truth.

$+39.45\%$ compared to DAF [10] in GTA $\rightarrow$ Cityscapes and about $+23.19\%$ compared to HRDA [11] in SYNTHIA $\rightarrow$ Cityscapes. This experiment demonstrated the effectiveness of our method in discriminating private classes while maintaining the performance of common classes. A more detailed examination revealed that we achieved a significant improvement in the private class IoU score to approximately $+40.79\%$ compared to the DAF [10], and also an increase in the common class mIoU score of about $+8.65\%$ compared to the HRDA [11]. This showed that our proposed method not only improved the performance of the private class but also contributed to a slight improvement in the common classes. This is because DECON loss encouraged features of the private class near the boundary to converge while distancing themselves from features of the common class. This reduced confusion between the common and private classes, improving predictions of the common class. Moreover, since OpenReMix was designed to learn size-invariant features regardless of the common and private classes, it enhanced the accuracy of predicting the shape of both common and private classes. We also compared with BUDA [48]. Since BUDA had access to predefined private category definitions and direct comparison was not practical, we offered a comparative analysis in supplementary material.

4.3. Qualitative Evaluation

To validate the performance of our method, we conducted additional qualitative evaluations to assess segmentation performance against baselines. We compared our method

Method				GTA5 → Cityscapes		
Config.	# Head	DECON	OpenReMix	Common	Private	H-Score
A	C			58.17	11.87	19.71
B	C+1			70.37	31.78	43.79
C	C+1	✓		71.16	48.34	57.57
D	C+1		✓	71.52	49.26	58.34
Ours	C+1	✓	✓	**72.47**	**55.42**	**62.81**

Table 2. Ablation study of the components in our BUS framework. Configuration A, B, C, and D represent confidence-based MIC, head-expansion, DECON loss, and OpenReMix, respectively.

with MIC, confidence-based MIC (Config. A), and the head-expansion approach (Config. B) in the GTA → Cityscapes (see Figure 3). Furthermore, we compared our method with the head-expansion approach (Config. B), the incorporation of a new DECON loss (Config. C), and the utilization of the new OpenReMix (Config. D) in the SYNTHIA → Cityscapes (see Figure 4). In Figure 3, we observed that the UDA method MIC, which was designed for UDA without considering unknown classes, struggled to detect the private classes in OSDA-SS. Even baselines like confidence-based MIC (Config. A) and head-expansion (Config. B) faced challenges in identifying private classes. Although head-expansion showed promise, it still had limitations in classifying specific pixels in private classes. In contrast, our method excelled, particularly in discerning object size. In Figure 4, our proposed DECON loss and OpenReMix yielded outstanding performance.

4.4. Ablation Study

Ablation Study. We conducted an ablation study for the proposed components of the BUS framework on GTA5 → Cityscapes. In Table 2, row A and B represented the confidence-threshold and head-expansion baselines, respectively. The confidence-threshold baseline (Config. A) recorded inferior performance compared to the head-expansion baseline (Config. B). It revealed that leveraging the expanded head was effective in detecting unknown classes, achieving H-Scores from 19.71% to 43.79%. When we combined DECON loss with the head-expansion, we achieved a +13.78% improvement in the H-Score (see row C). We also confirmed the effectiveness of our proposed OpenReMix. We gained a +14.55% improvement in the H-Score (see row D). Lastly, using both DECON and OpenReMix on head-expansion significantly improved H-Score of +19.02%. Figure 4 showed a clear improvement in predicting the unknown compared to the MIC with head-expansion approach (Config. B), and we observed synergy in overcoming individual drawbacks when compared to DECON loss (Config. C) and OpenReMix (Config. D).

GTA5 → Cityscapes			
# of Unknown	Config. A	Config. B	Ours
6	19.71	43.79	**62.81**
4	11.73	41.54	**54.72**
2	9.43	41.51	**56.82**

Table 3. The comparison of the number of unknown classes. Config. A denotes the confidence-based MIC and Config. B denotes the head-expansion baseline.

Unknown Proportion. We conducted the experiments under a various number of unknown classes. When the number of unknown classes was 6, 4, and 2, we compared our method with two MIC-based baselines. For the case of 4 unknown classes, we selected (“pole”, “traffic sign”, “person”, “rider”), and for the case of 2 unknown classes, we chose (“person”, “rider”). Table 3 showed that our proposed method consistently outperformed the baselines, regardless of the change in the number of unknown classes.

5. Conclusion

To tackle this challenging OSDA-SS task, we proposed a novel method named BUS. Our approach includes DECON loss, a new dilation-erosion-based contrastive loss designed to rectify less confident and erroneous predictions near class boundaries. In addition, we proposed OpenReMix guiding the model to acquire size-invariant features and efficiently train the expanded head by mixing unknown objects from the target into the source. Through extensive experiments, we demonstrated the efficacy of our proposed method on public benchmark datasets, surpassing existing approaches by a significant margin. We anticipate that our work will be widely applied in research or the industry field, providing a strong baseline to detect unexpected and unseen objects in mission-critical scenarios. As a limitation, our method is primarily based on pseudo-labeling. Therefore, if the model is poorly calibrated, it might not assign pixels belonging to the private classes as unknown. In this case, BUS might show a performance drop.

Acknowledgment

This work was supported by MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00258649) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation), and in part by the IITP grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and by the IITP grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 2012. 1

[2] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015. 1, 2

[4] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 2, 6

[6] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 6

[7] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 1, 3, 6

[8] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5, 6

[10] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6, 7

[11] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 6, 7

[12] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 5, 6

[13] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 5

[14] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 1, 5

[15] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2, 3, 6

[16] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[17] JoonHo Jang, Byeonghu Na, Dong Hyeok Shin, Mingi Ji, Kyungwoo Song, and Il-Chul Moon. Unknown-aware domain adversarial learning for open-set domain adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 6

[18] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 2, 6

[19] Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021. 2, 4

[20] Francesco Caliva, Claudia Iriondo, Alejandro Morales Martinez, Sharmila Majumdar, and Valentina Pedoia. Distance map loss penalty term for semantic segmentation. *arXiv preprint arXiv:1908.03679*, 2019.

[21] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 4

[22] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 2

[23] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2

[24] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[25] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 2

[26] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 2

[27] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019.

[28] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[29] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[30] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 3

[31] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020.

[32] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[34] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2019.

[35] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[36] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

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

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

[39] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[40] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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

[42] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[43] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

[44] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3, 4, 6

[45] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 3

[46] Qian Wang, Fanlin Meng, and Toby P Breckon. Progressively select and reject pseudo-labelled samples for open-set domain adaptation. *arXiv preprint arXiv:2110.12635*, 2021.

[47] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[48] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Handling new target classes in semantic segmentation with domain adaptation. *Computer Vision and Image Understanding*, 2021. 3, 7

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[50] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[51] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3, 5

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019.

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

[54] Li Gao, Jing Zhang, Lefei Zhang, and Dacheng Tao. Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3

[55] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems (NeurIPS)*, 2019.

[56] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

[57] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

[58] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[59] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *In Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3

[60] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. Deliberated domain bridging for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[61] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[62] Juwon Seo, Ji-Su Kang, and Gyeong-Moon Park. Lfs-gan: Lifelong few-shot image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[63] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems (NeurIPS)*, 2017. 4

[64] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020. 5

[65] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 5

[66] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[67] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009. 6

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[69] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 6

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 6