

Bidirectional Domain Mixup for Domain Adaptive Semantic Segmentation

Daehan Kim^{*1}, Minseok Seo^{*2}, Kwanyong Park³, Inkyu Shin³, Sanghyun Woo³, In-So Kweon³, Dong-Geol Choi^{†1}

¹ Hanbat National University, Korea

² SI Analytics, Korea

³ Korea Advanced Institute of Science and Technology (KAIST), Korea

daehan.kim@edu.hanbat.ac.kr, minseok.seo@si-analytics.ai, {pkyong7, dlsrbgg33, shwoo93, iskweon77}@kaist.ac.kr
dgchoi@hanbat.ac.kr

Abstract

Mixup provides interpolated training samples and allows the model to obtain smoother decision boundaries for better generalization. The idea can be naturally applied to the domain adaptation task, where we can mix the source and target samples to obtain domain-mixed samples for better adaptation. However, the extension of the idea from classification to segmentation (i.e., structured output) is nontrivial. This paper systematically studies the impact of mixup under the domain adaptive semantic segmentation task and presents a simple yet effective mixup strategy called Bidirectional Domain Mixup (BDM). In specific, we achieve domain mixup in two-step: cut and paste. Given the warm-up model trained from any adaptation techniques, we forward the source and target samples and perform a simple threshold-based cut out of the unconfident regions (**cut**). After then, we fill-in the dropped regions with the other domain region patches (**paste**). In doing so, we jointly consider class distribution, spatial structure, and pseudo label confidence. Based on our analysis, we found that BDM leaves domain transferable regions by **cutting**, balances the dataset-level class distribution while preserving natural scene context by **pasting**. We coupled our proposal with various state-of-the-art adaptation models and observe significant improvement consistently. We also provide extensive ablation experiments to empirically verify our main components of the framework. Visit our project page with the code at <https://sites.google.com/view/bidirectional-domain-mixup>.

Introduction

To reduce the annotation budget in semantic segmentation that require pixel level annotation, there have been many domain adaptation (DA) approaches using relatively inexpensive source (e.g. simulator-based) data (Richter et al. 2016; Ros et al. 2016) and unlabeled target (e.g. real) data. However, deep neural networks show poor generalization performance in real data because they are sensitive to domain misalignments such as layout (Li et al. 2020), texture (Yang and Soatto 2020), structure (Tsai et al. 2018), and class distribution (Zou et al. 2018b). To deal with it, many approaches have been proposed, including adversarial training (Tsai et al. 2018), entropy minimization (Vu et al. 2019),

and self-training (Zou et al. 2018b; Mei et al. 2020; Zhang et al. 2021).

Among them, cross-domain data mixing based approaches (Chen et al. 2021; Zhou et al. 2022a; Tranheden et al. 2021; Gao et al. 2021) recently show state-of-the-art performances. These methods composite the images of both domains and corresponding (pseudo) labels to generate domain-mixed training samples. Early works (Chen et al. 2021) are largely inspired by a popular data augmentation method, CutMix (Yun et al. 2019), and borrow some rectangular patches from one domain to fill-in the random hole of other domain of image. Many variants improve data mixing strategy with mixing the region of randomly sampled classes (Tranheden et al. 2021), heuristics on relationship between classes (Zhou et al. 2022a), and image-level soft mixup (Gao et al. 2021).

Motivated by the progress, we further delve into domain mixing approaches and propose **Bidirectional Domain Mixup (BDM)** framework. Beyond the previous unidirectional sample mixing (Chen et al. 2021; Zhou et al. 2022a; Tranheden et al. 2021; Gao et al. 2021), the framework mix the samples in both direction, mix the source patches on the target sample (i.e. source-to-target) and vice versa (i.e. target-to-source). Under the framework, we systematically study what makes good domain mixed samples to learn the transferable and generalized features for the target domain performance. Specifically, we mainly investigate two core steps of data mixing approach: 1) **Cut**: how can we identify uninformative patches and 2) **Paste**: which patches from other domains bring better supervision signals.

First, we promote to learn domain transferable and generalized features by cutting out the source-specific and noisily predicted region for source and target data, respectively. We found that the regions with low-confident model predictions is highly related to it, and cutting out such regions. To supplement scarce supervisory signal due to the cut process, we design the paste step to fulfill the three key functionalities: 1) As semantic segmentation network heavily rely on the context (Choi, Kim, and Choo 2020; Chen et al. 2017; Zheng et al. 2021; Yi et al. 2022), it is important to **maintain intrinsic spatial structure** of images. Thus we leverage spatial continuity to pick a patch that will be pasted on given the hole region. 2) Previous data mixing approaches (Zhou et al. 2022a; Tranheden et al. 2021) usually paste the randomly se-

^{*}These authors contributed equally.

[†]Corresponding authors: Dong-Geol Choi.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

lected classes with its correlated classes. This design choice exacerbates the class imbalanced problem (Gupta, Dollar, and Girshick 2019), resulting in low performance in sample-scarce class. Instead, we induce **class-balanced learning** by giving the high probability to paste the patches with rare classes. 3) Lastly, proposed mixing method **prevent the noisy learning** by avoiding to paste low-confident patches.

We combine these findings to achieve a new state-of-the-art in the standard benchmarks of domain adaptive semantic segmentation, GTA5 \rightarrow Cityscapes and SYNTHIA \rightarrow Cityscapes setting. In addition, BMD consistently provides significant performance improvements when build upon the various representative UDA approaches, including adversarial training (Tsai et al. 2018), entropy minimization (Vu et al. 2019), and self-training (Mei et al. 2020; Zhang et al. 2021).

Related Works

Unsupervised Domain Adaptive Semantic Segmentation. Unsupervised domain adaptation (UDA) aims to transfer the knowledge learned from the label-rich source domain to the unlabeled target domain. Early deep-learning-based approaches can be largely categorized into two groups. *Adversarial-based methods* are to align the feature distribution between the source and target domain via adversarial learning. Toward this goal, previous works investigate the three different levels: image level (Murez et al. 2018; Gong et al. 2019; Park et al. 2019; Chen et al. 2019), intermediate feature level (Hoffman et al. 2018; Liu, Zhang, and Wang 2021; Isobe et al. 2021), and output level (Tsai et al. 2018; Li, Yuan, and Vasconcelos 2019; Park et al. 2020). *Self-training-based methods* retrain the models on unlabeled target data with pseudo labels, which are generated based on the model’s predictions. Many works put research efforts into developing the advanced strategies for pseudo-label generation, such as image adaptive thresholding (Mei et al. 2020) and online pseudo-label generation (Zhang et al. 2021; Araslanov and Roth 2021). Recently, some works (Yang et al. 2020; Wang et al. 2020; Li et al. 2022) combine both types of approaches to achieve state-of-the-art performances.

Cross-domain Data Mixing Based Approaches. The cross-domain data mixing-based approach is one of the recent breakthroughs in domain adaptive semantic segmentation. These methods mix images from the different domains and utilize the mixed samples to train the generalized models. As an early attempt, DACS (Tranheden et al. 2021) borrows a popular data augmentation (Olsson et al. 2021) to mix the images, where the source patches of random classes are augmented to target images. Follow-up works devise additional regularization loss between original and mixed images by proposing regional contrastive consistency (Zhou et al. 2022b) or focusing on boundary regions (Liu et al. 2021). As another line, there are some trials to present new data mixing strategies. DSP (Gao et al. 2021) softly mixes the images from the source and target domain and uses the corresponding ratio to balance the loss of each domain. CAMix (Zhou et al. 2022a) considers contextual relationships in data mixing by relying on a prior class relationship.

Apart from most previous works, our framework generates the domain-mixed samples in a bidirectional way, source-to-target and target-to-source. Here we thoroughly study which factors make an effective domain-mixed sample in the context of domain-adaptive semantic segmentation.

Method

This section presents a bidirectional domain mixup framework for domain adaptive semantic segmentation. To bridge source and target domains that come from different distributions, we simulate intermediate domains by generating domain mixed samples. In the next section, we first describe the overall pipeline to train a network with domain-mixed samples. Next, we introduce how these samples are generated in detail.

The Bidirectional Domain Mixup Framework

Given a labeled source dataset $\mathcal{D}^S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$, and an unlabeled target dataset $\mathcal{D}^T = \{x_i^T\}_{i=1}^{N_T}$, our goal is to transfer the knowledge learned from source domain to unlabeled target domain. To do so, we utilize domain-mixed samples and propose a new data mixing method, bidirectional domain mixup (BDM), to generate them. As illustrated in Figure 1 and Alg 1, this framework comprises the following four major components.

- We first apply a previous domain adaptation method as a *warm-up*. Our framework allows any previous domain adaptation method. To show the generality, we choose the representative methods including adversarial training (Tsai et al. 2018), entropy minimization (Vu et al. 2019), and self-training (Mei et al. 2020; Zhang et al. 2021).
- *Domain-wise patch banks*, B^S and B^T , are constructed to store the samples of each domains. We divide images and corresponding pseudo labels into non-overlapping patches and the resulting pairs are stored in domain-wise patch banks. For the both domain, we adopt simple strategy (Yang and Soatto 2020) to generate pseudo labels, $\{\tilde{y}_i^S\}_{i=1}^{N_S}$ and $\{\tilde{y}_i^T\}_{i=1}^{N_T}$.
- During the training, a minibatch of source and target images, x^S and x^T , are sampled. *Bidirectional domain mixup (BDM)* generate domain-mixed samples, x_{mix}^S and x_{mix}^T , via mixture of images of one domain with patches from other domain. Thus, this cross-domain mixing is in bidirectional way, source-to-target and target-to-source. Specifically, some rectangular regions (*i.e.* patches) in target samples are cut and source patches retrieved from the source patch bank are pasted (*i.e.* source-to-target direction), and vice versa. We also apply same operation on labels, resulting in domain-mixed labels, y_{mix}^S and y_{mix}^T .
- Given the mixed images and labels, a segmentation network is trained with standard cross-entropy losses L_{cross} . The final loss is formulated as follows: $L_{final} = L_{cross}(x_{mix}^S, y_{mix}^S) + L_{cross}(x_{mix}^T, y_{mix}^T)$.

With the framework, we systematically study what makes good domain mixed samples to narrow the domain gap.

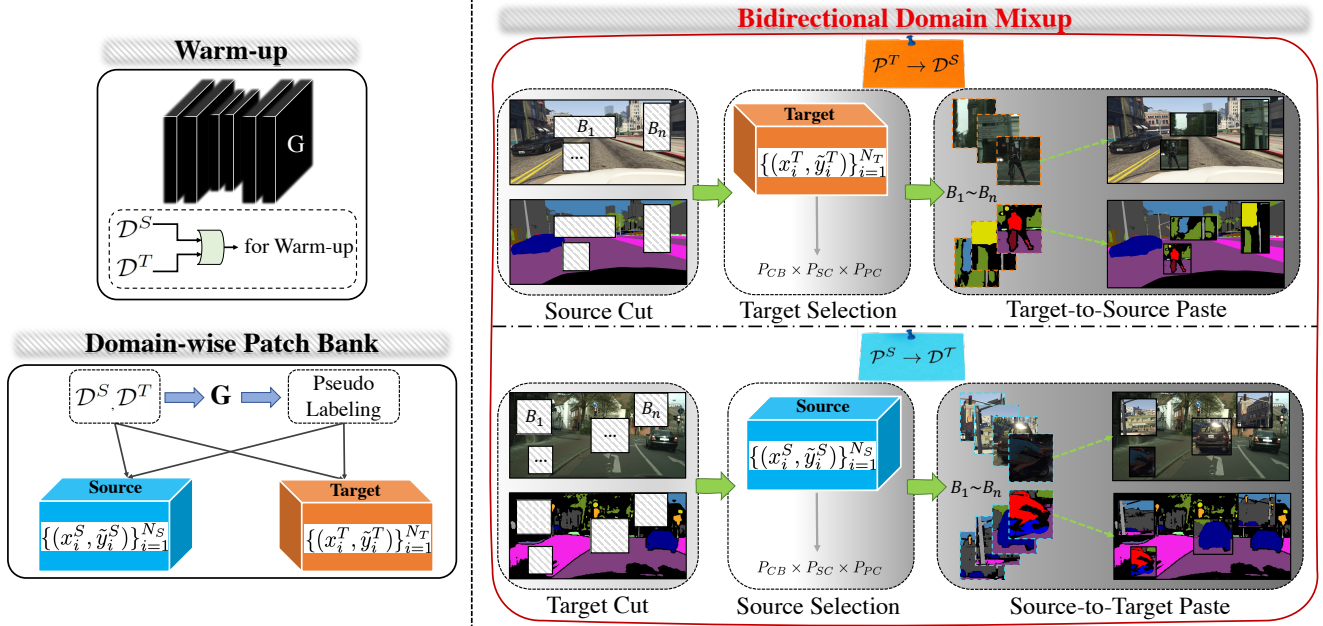


Figure 1: Overview of BDM framework. Before training a network with the proposed BDM framework, we first warm up the model with any previous UDA method. Given the model $G(\cdot)$, we generate pseudo labels for the source and target domain and store the images and corresponding pseudo labels in the domain-wise patch bank. During the training, the framework cut the patches of one domain sample and paste the selected patches from the cross-domain patch bank. The mixup process is conducted in a bidirectional way, target-to-source $P_T \rightarrow P_S$ and source-to-target $P_S \rightarrow P_T$. These samples are guided models to learn domain generalized features.

Algorithm 1: Bidirectional Domain Mixup

Input: warmup model G , labeled source $\mathcal{D}^S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$, unlabeled target $\mathcal{D}^T = \{x_i^T\}_{i=1}^{N_T}$, cutout γ in Eq. 2, horizontal W , Vertical H in Eq. 4

```

1: for  $i = 1$  to  $N_S, N_T$  do
2:    $\hat{y}_i^S, \hat{y}_i^T = \text{Pseudo}(G(x_i^S)), \text{Pseudo}(G(x_i^T))$ 
3:    $B^S, B^T = \text{Divide}(\{x_i^S, \hat{y}_i^S\}, \{x_i^T, \hat{y}_i^T\}, W, H)$ 
4: end for
5:  $\{\hat{x}_i^S, \hat{y}_i^S\} = \text{Cut}(\{x_i^S, \hat{y}_i^S\}, \gamma)$ 
6:  $\mathcal{P}_{selected}^T = \text{Selection}(B^T(P_{CB}, P_{SC}, P_{PC}))$ 
7:  $\{x_{mix}^S, y_{mix}^S\} = \text{Paste}(\{\hat{x}_i^S, \hat{y}_i^S\}, \mathcal{P}_{selected}^T)$ 
8:  $\{\hat{x}_i^T, \hat{y}_i^T\} = \text{Cut}(\{x_i^T, \hat{y}_i^T\}, \gamma)$ 
9:  $\mathcal{P}_{selected}^S = \text{Selection}(B^S(P_{CB}, P_{SC}, P_{PC}))$ 
10:  $\{x_{mix}^T, y_{mix}^T\} = \text{Paste}(\{\hat{x}_i^T, \hat{y}_i^T\}, \mathcal{P}_{selected}^S)$ 
11: return  $\{x_{mix}^S, y_{mix}^S\}, \{x_{mix}^T, y_{mix}^T\}$ 

```

Specifically, we mainly investigate two aspects: 1) how can we identify uninformative patches and 2) which patches from other domains bring good supervision signals for the target performance.

Cut

Cut is a process that masks out contiguous sections (*i.e.* multiple rectangular regions) of input and corresponding labels. We introduce widely used random patch cutout (DeVries

and Taylor 2017; Yun et al. 2019; Tranheden et al. 2021) and proposed confidence based cutout.

Random region cutout. Random region cutout is a simple but strong baseline that is adopted in various tasks such as UDA (Tranheden et al. 2021), Semi-DA (Chen et al. 2021) and SSL (Wang et al. 2022). When x and y are input, random region cutout could be formulated as follows:

$$x_{cut} = M \odot x, y_{cut} = M \odot y, \quad (1)$$

where $M \in \{0, 1\}^{W \times H}$ denotes a binary mask indicating where to drop out, and \odot is element-wise multiplication. Here, we zero out the multiple random regions of the mask to cut.

Confidence based cutout. The random region cut the regions regardless of whether it is informative. Instead, we target to cut where provide noisy supervision in terms of learning the generalized features. To this end, we see that it is important to discard the regions with low confident predictions for the following reasons: 1) for the source domain, it remove non-transferable and source specific regions (please refer to discussion section.), 2) it prevent to learn from noisy pseudo labels of target data.

Given the source images x_i^S and their pseudo labels \hat{y}_i^S , we calculates the ratio of the uncertain region over the randomly generated region $\mathbf{B} = (r_x, r_y, r_w, r_h)$. Here we see the ignored region of pseudo labels as the uncertain region. If the ratio of the uncertain region is above the cutout thresh-

old γ , the region is cut. The proposed cutout is summarized as follows:

$$\hat{x}_i^S, \hat{y}_i^S = \begin{cases} \text{Cutout}(x_i^S, \tilde{y}_i^S), & \text{if } \mathcal{H}(\tilde{y}_i^S, \mathbf{B}) > \gamma, \\ x_i^S, \tilde{y}_i^S & \text{otherwise,} \end{cases} \quad (2)$$

where \mathcal{H} denote a function that computes the ratio of the uncertain region over region \mathbf{B} . And, the threshold γ is set to 0.2. Similar to Eq. 2, low confident regions of target samples are also masked.

Paste

To generate the domain-mixed samples, we sample the patch from the other domain and paste it back to the region that are cutout. This enables the resulting samples to include both source and target patches. While already effective (see Table 4), we see that there are more room for the improvement.

We additionally consider three important factors during the patch sampling. First, we introduce **class-balanced** sampling. As random sampling tend to bias the mixup samples mainly toward the frequent classes, we offset this undesirable effect using class-balanced sampling. Since the long-tailed category distribution tends to be similar for different domains, we compute the class distribution of the source domain (Zhang et al. 2021; Li et al. 2022) to provide more chance for patches that include rare classes. Second, we consider **spatial continuity** of the natural scene. The random sampling can produce mixed samples that are geometrically unnatural, and thus causes severe train and test time inconsistency. Instead, based on the fact that the vehicle ego-centric view has strong fixed spatial priors (Choi, Kim, and Choo 2020; Zou et al. 2018a), we compute spatial priors for each semantic class (Zou et al. 2018a) in the source domain and use this statistics during sampling. In this way, we successfully synthesize geometrically reasonable mixed samples (see Figure 2). Finally, we take **pseudo label confidence** into account. The rationale behind this strategy is to increase the sampling probability of the patches that include confident pixels (and vice versa). This is also strongly connected to the cut operation, where we dropped the regions that are not domain transferable, while here we perform the opposite. We observe that all these components impact the final performance, and the best mixup results are achieved when these are used jointly.

Patch Generation. In the source and target datasets, classes such as *train* and *bike* have a small number of samples, so to consider the class online, it is necessary to find a sample including a 6 rare class from all samples. Since this method is memory inefficient, we cut the patch containing each class and save it.

We generate a patch by cutting the pseudo labeled datasets $\{(x_i^S, \tilde{y}_i^S)\}_{i=1}^{N_S}$ and $\{(x_i^T, \tilde{y}_i^T)\}_{i=1}^{N_T}$ at equal intervals by the number of horizontal \mathbf{W} and vertical \mathbf{H} . Therefore, one image and its pseudo label are divided into $\mathbf{W} \times \mathbf{H}$ patches. After that, the patches extracted from each location are grouped into one patch sequence. Next, the patch containing the class $\{C_i\}_{i=1}^K$, \mathbf{K} is the number of classes, from the pseudo label of the patch sequence representing each spatial location is

stored in the child patch sequence, and it is possible to duplicate it. Therefore, patches considering spatial location and class existence are grouped into a total of $\mathbf{W} \times \mathbf{H} \times \mathbf{K}$ patch sequence. Finally, at the patch sequence of $\mathbf{W} \times \mathbf{H} \times \mathbf{K}$, the normalized confidence calculated by the method in Eq. 5 is sorted in ascending order, and \mathbf{R} patch sequences are generated at regular intervals. The number of finally generated patch sequence is $\mathbf{W} \times \mathbf{H} \times \mathbf{K} \times \mathbf{R}$

Patch Selection

Class-balanced Patch Sampling. In addition to Zipfian distribution of object categories (Manning and Schutze 1999), the difference in the intrinsic size of each object makes pixel-level class imbalances more severe in semantic segmentation. However, the random patch sampling for paste make bias toward the frequent classes, as it naturally follow the probability of existence.

To alleviate the class imbalance problem in paste process, we propose patch-level oversampling. Given the total number of pixels for each classes in the source labels $\{\bar{N}^i\}_{i=1}^K$, the probability P_{CB} that each class patch is selected can be formulated as follows:

$$\hat{P}_{CB} = \left\{ \left(-\log \left(\frac{\bar{N}^i}{\sum_{i=0}^{\mathbf{K}} \bar{N}^i} \right) \right)^\alpha \right\}_{i=1}^K, \quad (3)$$

$$P_{CB} = \left\{ \left(\frac{\hat{P}_{CB}^i}{\sum_{i=0}^{\mathbf{K}} \hat{P}_{CB}^i} \right) \right\}_{i=1}^K,$$

where \mathbf{K} is the number of classes, α is the sharpening coefficient. We set α to 2 in all experiments.

Sampling with Spatial Continuity. The spatial layout between the source (synthetic) and target dataset share large similarities. For example, there are common rules like cars can't fly in the sky (Choi, Kim, and Choo 2020) in both the domains. Therefore, we propose spatial continuity based paste that considers spatial relationship instead of pasting patches at random positions. The probability of selecting each location $\{\text{Patches}_i\}_{i=1}^{\mathbf{W} \times \mathbf{H}}$ of the patches generated through Patch Generation section to be mixed with the patch locations cutout through Cut section is calculated as follows:

$$\begin{aligned} \hat{S}C &= \text{argmax} \{ \text{SC}_i(o_w, o_h)_{i=1}^K \}, \\ P_{SC} &= \{ \hat{S}C(\text{Patches}_i(o_{\hat{w}^i}, o_{\hat{h}^i})) \}_{i=1}^{\mathbf{W} \times \mathbf{H}}, \end{aligned} \quad (4)$$

where $\{\text{SC}_i\}_{i=1}^{\mathbf{K}}$ is the source domain class-wise spatial prior kernel map generated by CBST-SP (Zou et al. 2018a), where o_h, o_w is the center coordinates of the cutout patch, o_h^i, o_w^i is the center coordinates of the patch at the i -th location. Note that, we normalized the sum of the set P_{SC} to 1.

Sampling with Normalized Confidence. Opposite to confidence based cutout, in the paste, we give high probability to the patches that include confident pixels. To faithfully measure the confidence level of a patch, we take the difficulty of each class into account and design the normalized

Method	Head Classes														Tail Classes						
	mIoU	mIoU-tail	Road	Sidewalk	Building	Wall	Fence	Pole	Vegetation	Terrain	Sky	Person	Car	Truck	Bus	Light	Sign	Rider	Train	Motorcycle	Bike
Source Only	36.6	24.0	75.8	16.8	77.2	12.5	21.0	25.5	81.3	24.6	70.3	53.8	49.9	17.2	25.9	30.1	20.1	26.4	6.5	25.3	36.0
Adaptseg (Tsai et al. 2018)	41.4	25.0	86.5	25.9	79.8	22.1	20.0	23.6	81.8	25.9	75.9	57.3	76.3	29.8	32.1	33.1	21.8	26.2	7.2	29.5	32.5
ADVENT (Vu et al. 2019)	45.5	25.7	89.4	33.1	81.0	26.6	26.8	27.2	83.9	36.7	78.8	58.7	84.8	38.5	44.5	33.5	24.7	30.5	1.7	31.6	32.4
CCM (Li et al. 2020)	49.9	26.9	93.5	57.6	84.6	39.3	24.1	25.2	85.0	40.6	86.5	58.7	85.8	49.0	56.4	35.0	17.3	28.7	5.4	31.9	43.2
IAST (Mei et al. 2020)	51.5	34.0	93.8	57.8	85.1	39.5	26.7	26.2	84.9	32.9	88.0	62.6	87.3	39.2	49.6	43.1	34.7	29.0	23.2	34.7	39.6
DACS (Tranheden et al. 2021)	52.1	32.6	89.9	39.6	87.8	30.7	39.5	38.5	87.9	43.9	88.7	67.2	84.4	45.7	50.1	46.4	52.7	35.7	0.0	27.2	33.9
DSP (Gao et al. 2021)	55.0	36.9	92.4	48.0	87.4	33.4	35.1	36.4	87.7	43.2	89.8	66.6	89.9	57.1	56.1	41.6	46.0	32.1	0.0	44.1	57.8
CAMix (Zhou et al. 2022a)	55.2	37.9	93.3	58.2	86.5	36.8	31.5	36.4	87.2	44.6	88.1	65.0	89.7	46.9	56.8	35.0	43.5	24.7	27.5	41.1	56.0
CorDA (Wang et al. 2021)	56.6	38.5	94.7	63.1	87.6	30.7	40.6	40.2	87.6	47.0	89.7	66.7	90.2	48.9	57.5	47.8	51.6	35.9	0.0	39.7	56.0
ProDA (Zhang et al. 2021)	57.5	42.0	87.8	56.0	79.7	46.3	44.8	45.6	88.6	45.2	82.1	70.7	88.8	45.5	59.4	53.5	53.5	39.2	1.0	48.9	56.4
DAP (Huo et al. 2022)	59.8	44.6	94.5	63.1	89.1	29.8	47.5	50.4	89.5	50.2	87.0	73.6	91.3	50.2	52.9	56.7	58.7	38.6	0.0	50.2	63.5
Adaptseg (Tsai et al. 2018) + Ours	57.4	44.4	89.3	50.0	88.4	45.6	45.4	41.1	78.0	35.6	82.3	69.2	87.5	55.7	57.8	49.9	60.2	45.6	8.1	45.5	57.2
ADVENT (Vu et al. 2019) + Ours	57.6	38.9	91.3	51.8	86.7	49.9	49.2	53.3	85.8	47.9	85.7	62.3	87.8	55.5	54.4	43.1	43.3	45.9	4.4	46.3	50.4
IAST (Mei et al. 2020) + Ours	61.0	46.5	92.1	59.6	89.9	52.9	55.7	49.2	89.3	46.7	86.3	59.1	88.3	54.8	55.9	44.6	45.8	42.0	39.2	50.3	57.3
ProDA (Zhang et al. 2021) + Ours	63.9	47.8	89.2	60.1	83.8	61.5	63.6	66.7	90.4	51.1	83.5	72.6	88.0	51.2	65.3	58.2	59.3	47.8	1.0	60.1	60.9
Target Only	64.5	53.0	96.2	75.5	87.7	38.0	39.6	43.4	88.2	52.4	89.5	69.7	91.4	66.2	69.7	46.6	62.8	49.5	45.0	49.0	65.1

Table 1: Comparison with state-of-the-art models on GTA5 \rightarrow Cityscapes. We highlight the mIoU of tail classes (*i.e.* mIoU-tail) along with per-class IoU and overall mIoU. Our results are averaged over five runs.

confidence of a patch. We first calculate the average confidence of classes using the set of pseudo labels and use it to represent the difficulty of each class. Then, the confidence of patches in the patch bank is normalized at pixel-level by subtracting the difficulty score according to predicted classes (Norm). Intuitively, it measure the *relative* confidence level. The resulting normalized confidence maps are spatially averaged (Average), and patches are sorted in ascending order according to it (Sort).

$$\begin{aligned}
\hat{B}^T &= \{Average(Norm(B_i^T))\}_{i=1}^{N_T}, \\
\hat{B}^S &= \{Average(Norm(B_i^S))\}_{i=1}^{N_S}, \\
\bar{B}^T &= Sort(\hat{B}^T), \\
\bar{B}^S &= Sort(\hat{B}^S)
\end{aligned} \tag{5}$$

Finally, the patch is divided into three batches: the low, the middle, and the high confidence group. The probability P_{PC} that the patch in each group is selected is $\{0.1, 0.3, 0.6\}$.

Probability of selection of each patch sequence. We select the patch jointly considering class balance, spatial continuity, and pseudo label confidence for BDM. Since each probability is independent, the probability that each patch sequence is selected is $P_{CB} \times P_{SC} \times P_{PC}$.

Experiments

In this section, we present experimental results to validate the proposed BDM for domain adaptive semantic segmentation.

We first describe experimental configurations in detail. After that, we validate our BDM on two public benchmark datasets, GTA5 and SYNTHIA (Ros et al. 2016) datasets, and provide detailed analyses. Note that the Intersection-over-Union (IoU) metric is used for all the experiments.

Experimental Settings

Dataset. We evaluate our proposed Bidirectional Domain Mixup on two popular domain adaptive semantic segmentation benchmarks (SYNTHIA \rightarrow Cityscapes, and GTA5 \rightarrow Cityscapes). Cityscapes (Cordts et al. 2016) is a real-world urban scene dataset consisting of a training set with 2,975 images, a validation set with 500 images and a testing set with 1,525 images. We use the unlabeled training dataset as $\{D_i^T\}_{i=1}^{2,975}$ and evaluate our Bidirectional Domain Mixup with 500 images from the validation set. SYNTHIA (Ros et al. 2016) is a synthetic urban scene dataset. We pick SYNTHIA-RAND-CITYSCAPES subset as the source domain, which shares 16 semantic classes with Cityscapes. In total, 9,400 images from SYNTHIA dataset are used as source domain training data $\{D_i^S\}_{i=1}^{9,400}$ for the task. GTA5 (Richter et al. 2016) dataset is another synthetic dataset sharing 19 semantic classes with Cityscapes. 24,966 urban scene images are collected from a physically-based rendered video game Grand Theft Auto V (GTAV) and are used as source training data $\{D_i^S\}_{i=1}^{24,966}$. We view 6 and 5 with relatively few training samples as tail-classes for each source domain (GTA5, SYNTHIA), respectively.

Method	Head Classes										Tail Classes				
	mIoU	mIoU-tail	Road	Sidewalk	Building	Vegetation	Sky	Person	Car	Bus	Light	Sign	Rider	Motorcycle	Bike
Source Only	40.3	20.8	64.3	21.3	73.1	63.1	67.6	42.2	73.1	15.3	7.0	27.7	19.9	10.5	38.9
Adaptseg (Tsai et al. 2018)	45.8	18.5	79.5	37.1	78.2	78.0	80.3	53.7	67.1	29.4	9.3	10.6	19.2	21.8	31.6
ADVENT (Li et al. 2020)	48.0	16.9	85.6	42.2	79.7	80.4	84.1	57.9	73.3	36.4	5.4	8.1	23.8	14.2	33.0
CCM (Li et al. 2020)	52.9	29.0	79.6	36.4	80.6	81.8	77.4	56.8	80.7	45.2	22.4	14.9	25.9	29.9	52.0
IAST (Mei et al. 2020)	57.0	35.2	81.9	41.5	83.3	83.4	85.0	65.5	86.5	38.2	30.9	28.8	30.8	33.1	52.7
DACS (Tranheden et al. 2021)	54.8	32.1	80.5	25.1	81.9	83.6	90.7	67.6	82.9	38.9	22.6	23.9	38.3	28.4	47.5
DSP (Gao et al. 2021)	59.9	35.9	86.4	42.0	82.0	87.2	88.5	64.1	83.8	65.4	31.6	33.2	31.9	28.8	54.0
CAMix (Zhou et al. 2022a)	59.7	33.8	91.8	54.9	83.6	83.8	87.1	65.0	85.5	55.1	23.0	29.0	26.4	36.8	54.1
CorDA (Wang et al. 2021)	62.8	41.5	93.3	61.6	85.3	84.9	90.4	69.7	85.6	38.4	36.6	42.8	41.8	32.6	53.9
ProDA (Zhang et al. 2021)	62.0	40.3	87.8	45.7	84.6	88.1	84.4	74.2	88.2	51.1	54.6	37.0	24.3	40.5	45.6
DAP (Huo et al. 2022)	64.3	48.7	84.2	46.5	82.5	89.3	87.5	75.7	91.7	73.5	53.6	45.7	34.6	49.4	60.5
Adaptseg* (Tsai et al. 2018) + Ours	50.6	24.1	84.5	44.3	79.5	84.2	83.3	60.1	69.3	33.2	19.6	18.7	23.6	24.0	34.7
ADVENT (Tsai et al. 2018) + Ours	53.8	27.1	85.4	47.7	82.9	86.5	85.7	64.5	72.0	39.3	25.2	22.2	25.6	26.1	36.4
IAST (Mei et al. 2020) + Ours	62.9	42.6	88.2	50.2	88.5	85.6	89.7	70.2	88.3	44.8	42.3	33.5	39.0	39.4	59.0
ProDA (Zhang et al. 2021) + Ours	66.8	47.7	91.0	55.8	86.9	85.8	85.7	84.1	86.0	55.2	58.3	44.7	40.3	45.0	50.6
Target Only	72.3	54.6	96.2	75.5	87.7	88.2	89.5	69.7	91.4	69.7	46.6	62.8	49.5	49.0	65.1

* Note that pre-trained weights are not provided, so use them after reproduction.

Table 2: Comparison with state-of-the-art models on SYNTHIA \rightarrow Cityscapes. Our results are averaged over five runs.

Training. In all our experiments, we used Deeplabv2 (Chen et al. 2017) architecture with pre-trained ResNet101 (He et al. 2016) as backbone on ImageNet (Deng et al. 2009) and MSCOCO (Lin et al. 2014), and for fair performance evaluation, all hyperparameter settings such as batch size, learning rate, and iteration follow the standard protocol (Tsai et al. 2018).

To show the generality of BDM, we choose multiple representative UDA approaches as warm-up method, including Adaptseg (Tsai et al. 2018) (adversarial training), ADVENT (Vu et al. 2019) (entropy minimization), IAST (Mei et al. 2020) (self-training), and ProDA (Zhang et al. 2021) (self-training) as warm-up models, respectively. We use official implementation of them¹²³⁴. Note that all ablation studies were performed with Adaptseg.

Given the warm-up model, we further train the model with proposed BDM framework for 1,000,000 iterations. For the stable training, EMA framework (Olsson et al. 2021) is adopted. Unless otherwise specified, we set the number of patches \mathbf{W} and \mathbf{H} as 4 and 3. For the number of randomly generated boxes in cut process, we choose 4 as default. The number of classes \mathbf{K} is 19 and 13 when we use GTA5 and SYNTHIA as the source domain, respectively. Finally, the number of pseudo-label reliability intervals, \mathbf{R} , was set to 3 in all experiments.

¹<https://github.com/wasidennis/AdaptSegNet>

²<https://github.com/Raykoooo/IAST>

³<https://github.com/microsoft/ProDA>

⁴<https://github.com/valeoai/ADVENT>

Comparison with State-of-the art

In this section, we compare our proposed method with the top-performing UDA approach.

Table 1 shows the comparisons on GTA5 \rightarrow Cityscapes setting. DACS, which classmix at random locations without considering the class distribution of the source dataset, significantly improved performance in classes with a large number of samples, but significantly decreased in tail classes such as *bike* and *train*. CAMix, a classmix method considering the contextual relationship, solved the problem of performance degradation of tail classes through consistency loss and dynamic threshold. However, considering only the contextual relationship, the performance decreased in *wall*, *light*, and *rider* classes, which have significantly different contextual relationship between GTA5 and Cityscapes.

On the other hand, our BDM jointly consider class balance, spatial continuity, and pseudo label confidence and achieve state-of-the-art with an mIoU score of 63.9% when ProDA was selected as the warm-up model. Despite the great overall scores, it showed a low IoU in the *train* class. The rationale behind this is the severely poor performance of the chosen warm-up model in that class. Instead, when the warm-up model is switched to IAST, we achieved much improved scores (23.2% \rightarrow 39.2%) IoU in the *train* class.

Last but not least, our method shows consistent performance improvement with four different warm-up models, showing the generality of our framework.

Table 2 shows the comparisons of SYNTHIA \rightarrow Cityscapes adaptation. Again, our BDM achieved state-of-

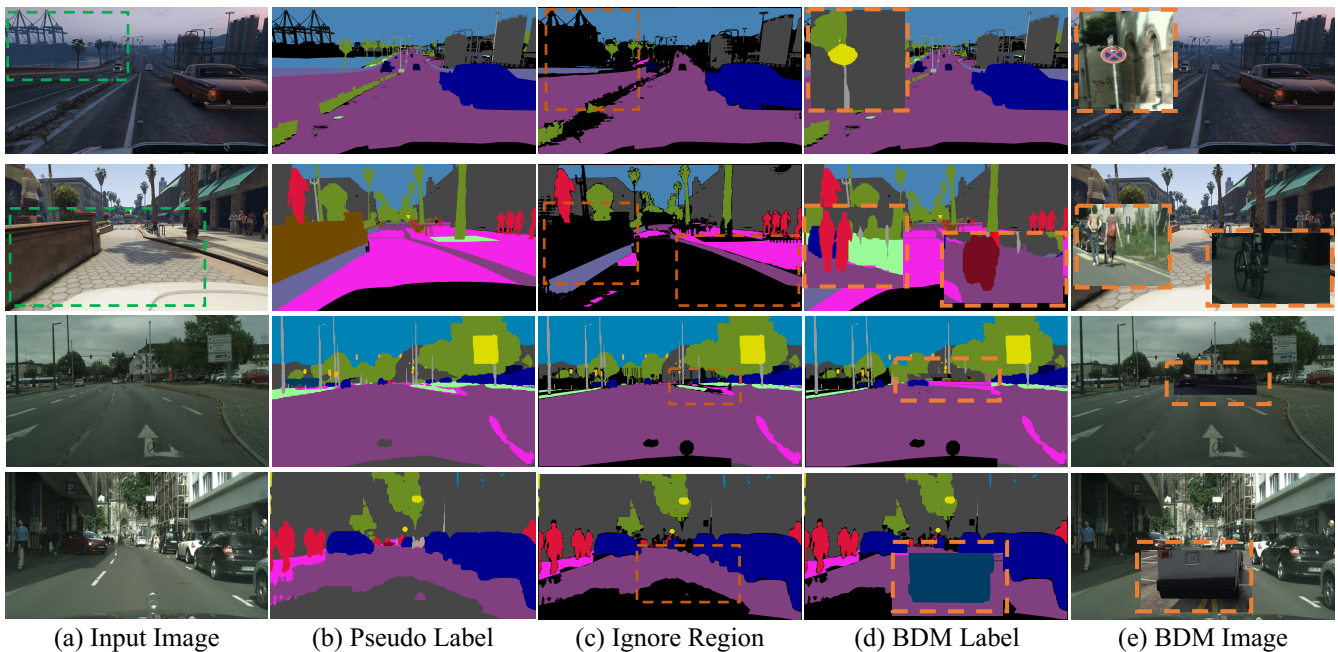


Figure 2: Qualitative results of cross domain mixed samples generated with the proposed BDM. Two rows from top to bottom: target-to-source, source-to-target. green box has a high probability of being Cut because ignore pixels are dense.

Method	BD.	Cut	Paste	mIoU	gain
Warm-up (Tsai et al. 2018)	-	-	-	46.4	+0
DACS (Tranheden et al. 2021)	-	Rand	Rand	50.9	+9.5
	✓	Rand	Rand	53.6	+12.2
Ours	✓	Ours	Rand	53.8	+12.4
	✓	Rand	Ours	56.8	+15.4
	✓	Ours	Ours	57.4	+16.0
Ours-T2S	-	Ours	Ours	56.0	+14.6
Ours-S2T	-	Ours	Ours	54.8	+13.4

Table 3: Component-wise Ablation Study.

the-art with an mIoU score of 66.8% when ProDA was selected as the warm-up model. These experimental results indicate that BDM is valid not only in the GTA5 dataset, where the scene layout between the source and target dataset is highly similar but also in the SYNTHIA dataset which includes images with different viewpoints (e.g. top-down view).

Ablation Study

Component-wise Ablation. First, we explore the importance of the bidirectional (BD.) mixup (please refer to Table 3). We implement two unidirectional variants, Ours-T2S and Ours-S2T, which mix the samples in target-to-source and source-to-target, respectively. As degraded results indicate, both directions of the mixture are important, showing the effectiveness of our BDM framework. We also report the results of DACS (Tranheden et al. 2021) and its bidirectional extension. For a fair comparison, we re-implement their mixing strategy on our framework. Again, the design

Method	CB	SC	PC	mIoU	gain
Ours _{cut}	-	-	-	48.9	+0
Ours	-	-	-	53.8	+4.9
	✓	-	-	56.1	+7.2
	✓	✓	-	56.4	+7.5
	✓	-	✓	57.0	+8.1
	✓	✓	✓	57.4	+8.5

Table 4: Ablation study on Paste.

of bidirectional mixing improves the performance by 2.7 mIoU.

We also investigate the impact of the cut and paste mechanism. Both proposed mechanism outperforms the popular random baseline. In particular, when we apply a random baseline for both the cut and paste, the model degenerate into a bidirectional extension of DACS (Tranheden et al. 2021). Compared to it, our final model boosts the performance by 3.8 mIoU.

Ablation Study on Paste. Table 4 study the effects of class balance (CB), spatial continuity (SC), and pseudo label confidence (PC) on the paste mechanism. Ours_{cut} refers to the model trained with only the proposed cut process. As shown in Table 4, the CB component showed the largest performance improvement. Also, SC and PC showed 0.3 mIoU and 0.9 mIoU additional improvement, respectively. In addition, the best performance was achieved when all three components were used. These experimental results indicate that all components of CB, SC and PC are complementary in BDM.

	No Cut	Cut Top 50% Conf.	Cut Bottom 50% Conf.
mIoU	36.6	27.8 (-8.8)	41.3 (+4.7)

Table 5: Analysis on source cut region.

Model	Recall	Precision
Warm-up (Tsai et al. 2018)	97.6	73.2

Table 6: Analysis on Low-confident Regions.

Discussion

Connection between Low-confident and Transferable Region on Source Domain. We aim to cutout the non-transferable regions of the source domain for learning generalized features. To realize it, we remove the patches with low confident prediction. Here, we validate the design by showing that low-confident source regions are highly correlated with non-transferable regions. To this end, we first train multiple source-only models with varying regions of cut and measure the performance on the target dataset (see Table 5). Compared to the default model that uses all the source labels (*i.e.* No Cut in the Table 5), the performance is notably increased by 4.7 mIoU when we discard the labels of the bottom 50% of low confident pixels. On the contrary, cutting the high confident region makes significant performance degradation. It implies that training models with low confident pixels promote overfitting to the source domain.

Moreover, we examine how the selected low-confident regions correlates with lower domain transferrability. Based on the assumption that the target-only model has oracle feature representations, we compare the overlapping regions between the selected low-confident regions from our strategy and the target-only model. The results are summarized in Table 6. We see the regions highly overlap, implying that the simple threshold-based cutout can already leave the highly transferrable regions, and help better adaptation of the model.

Qualitative analysis of Bidirectional Domain Mixup.

Figure 2 is a sample of training images generated through BDM. BDM receives Figure 2-(c) and cuts the ignore region, and then pastes another patch to the corresponding region as shown in Figure 2-(d). Our framework is trained with Figure 2-(e) and Figure 2-(d) as pairs.

Figure 3-(a) shows the number of pixel-level supervision by class with and without patch level oversampling considering CB. As shown in the figure, if patch level oversample is applied (*e.g.* in GTA5, the probability that *bike* will be selected is 23%), it can be seen that the imbalance problem of pixel-level supervision is greatly alleviated. Figure 3-(b) shows the spatial continuity of random cutmix and BDM. As shown in the figure, our proposed BDM can confirm that spatial continuity is maintained.

Limitations of Bidirectional Domain Mixup. Since BDM uses pseudo labels of the source and target dataset, the performance of the warm-up model greatly affects per-

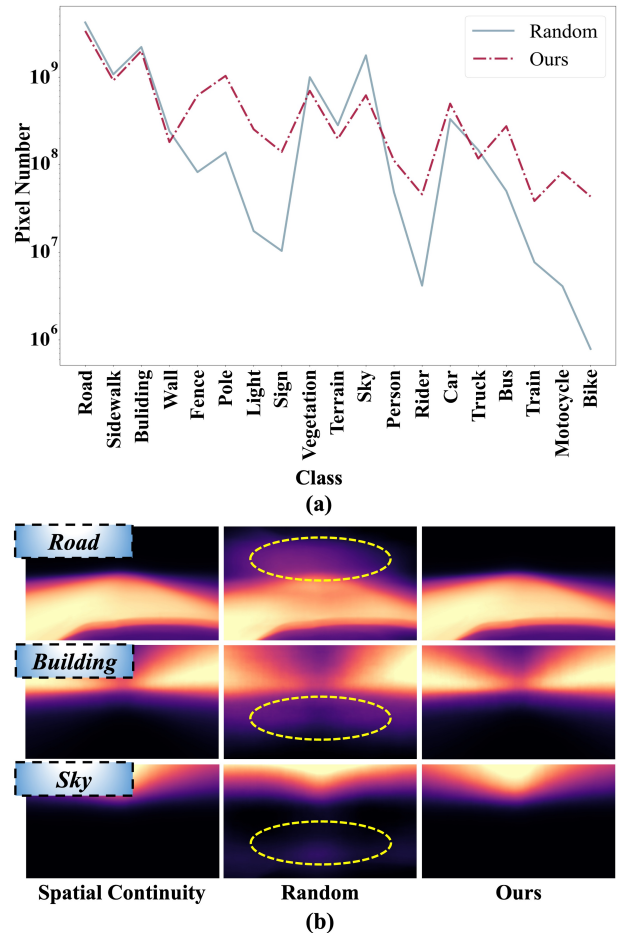


Figure 3: Visualization of class distribution and spatial continuity kernel in BDM. (a) shows comparison on the imbalance ratio of class-wise supervision. For each classes, we report the number of supervisory pixels that used in the training. Our method provide more class-balanced learning signal. (b) is the result of comparing the random cutmix and ours with the spatial continuity kernel of BDM.

formance of the final model. For example, as shown in Table 1, IoU and mIoU differ greatly in the warm-up model. However, it showed consistent performance improvement in representative UDA methods, such as entropy minimization (Vu et al. 2019), adversarial training (Tsai et al. 2018), and self-training (Mei et al. 2020; Zhang et al. 2021).

Conclusions

In this paper, we proposed Bidirectional Domain Mixup (BDM), a cutmix method that cut the low confidence region and selects a patch to paste according to class-balance(CB), spatial continuity(SC), and pseudo label confidence(PC) in the corresponding region. We observed that there are transferable and non-transferable regions in the source dataset, and showed that the non-transferable region can be easily removed through the pseudo label of the warm-up model. Finally, we have shown through extensive experiments that

CB, SC, and PC are important components for the paste mechanism in the context of UDA. Our proposed BDM achieves state-of-the-art in GTA5 to cityscapes benchmark and SYNTHIA to cityscapes benchmark with a large gap. We hope that our research will be applied not only to UDA but also to related fields, such as semi-supervised domain adaptation, and semi-supervised segmentation fields.

Acknowledgements

This research was supported by the the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2022R1F1A1075019, No. 2021M3E8A2100446).

References

- Araşlanov, N.; and Roth, S. 2021. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15384–15394.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.
- Chen, S.; Jia, X.; He, J.; Shi, Y.; and Liu, J. 2021. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11018–11027.
- Chen, Y.-C.; Lin, Y.-Y.; Yang, M.-H.; and Huang, J.-B. 2019. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1791–1800.
- Choi, S.; Kim, J. T.; and Choo, J. 2020. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9373–9383.
- Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
- Gao, L.; Zhang, J.; Zhang, L.; and Tao, D. 2021. Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2825–2833.
- Gong, R.; Li, W.; Chen, Y.; and Gool, L. V. 2019. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2477–2486.
- Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, 1989–1998. Pmlr.
- Huo, X.; Xie, L.; Hu, H.; Zhou, W.; Li, H.; and Tian, Q. 2022. Domain-Agnostic Prior for Transfer Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7075–7085.
- Isobe, T.; Jia, X.; Chen, S.; He, J.; Shi, Y.; Liu, J.; Lu, H.; and Wang, S. 2021. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8187–8196.
- Li, G.; Kang, G.; Liu, W.; Wei, Y.; and Yang, Y. 2020. Content-consistent matching for domain adaptive semantic segmentation. In *European conference on computer vision*, 440–456. Springer.
- Li, R.; Li, S.; He, C.; Zhang, Y.; Jia, X.; and Zhang, L. 2022. Class-Balanced Pixel-Level Self-Labeling for Domain Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11593–11603.
- Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6936–6945.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Liu, Y.; Deng, J.; Gao, X.; Li, W.; and Duan, L. 2021. BAPA-Net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8801–8811.
- Liu, Y.; Zhang, W.; and Wang, J. 2021. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1215–1224.
- Manning, C.; and Schütze, H. 1999. *Foundations of statistical natural language processing*. MIT press.
- Mei, K.; Zhu, C.; Zou, J.; and Zhang, S. 2020. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, 415–430. Springer.

- Murez, Z.; Kolouri, S.; Kriegman, D.; Ramamoorthi, R.; and Kim, K. 2018. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4500–4509.
- Olsson, V.; Tranheden, W.; Pinto, J.; and Svensson, L. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1369–1378.
- Park, K.; Woo, S.; Kim, D.; Cho, D.; and Kweon, I. S. 2019. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1248–1257.
- Park, K.; Woo, S.; Shin, I.; and Kweon, I. S. 2020. Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 10869–10880.
- Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*, 102–118. Springer.
- Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. 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*, 3234–3243.
- Tranheden, W.; Olsson, V.; Pinto, J.; and Svensson, L. 2021. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1379–1389.
- Tsai, Y.-H.; Hung, W.-C.; Schuler, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7472–7481.
- Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2517–2526.
- Wang, H.; Shen, T.; Zhang, W.; Duan, L.-Y.; and Mei, T. 2020. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, 642–659. Springer.
- Wang, Q.; Dai, D.; Hoyer, L.; Van Gool, L.; and Fink, O. 2021. Domain Adaptive Semantic Segmentation with Self-Supervised Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. *arXiv preprint arXiv:2203.03884*.
- Yang, J.; An, W.; Wang, S.; Zhu, X.; Yan, C.; and Huang, J. 2020. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European conference on computer vision*, 480–498. Springer.
- Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.
- Yi, J. S. K.; Seo, M.; Park, J.; and Choi, D.-G. 2022. Using Self-Supervised Pretext Tasks for Active Learning. In *Proc. ECCV*.
- Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.
- Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021. 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*, 12414–12424.
- Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.
- Zhou, Q.; Feng, Z.; Gu, Q.; Pang, J.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2022a. Context-aware mixup for domain adaptive semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Zhou, Q.; Zhuang, C.; Yi, R.; Lu, X.; and Ma, L. 2022b. Domain Adaptive Semantic Segmentation via Regional Contrastive Consistency Regularization. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06. IEEE.
- Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018a. Domain adaptation for semantic segmentation via class-balanced self-training. *arXiv preprint arXiv:1810.07911*.
- Zou, Y.; Yu, Z.; Kumar, B. V.; and Wang, J. 2018b. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *Proceedings of the European Conference on Computer Vision (ECCV)*.