Contrast, Stylize and Adapt: Unsupervised Contrastive Learning Framework for Domain Adaptive Semantic Segmentation

Anonymous CVPR submission

Paper ID 24

Abstract

*To overcome the domain gap between synthetic and real-world datasets, unsupervised domain adaptation methods have been proposed for semantic segmentation. Majority of the previous approaches have attempted to reduce the gap either at the pixel or feature level, disregarding the fact that the two components interact positively. To address this, we present **CON**trastive **FE**a**T**ure and p**I**xel alignment (CONFETI) for bridging the domain gap at both the pixel and feature levels using a unique contrastive formulation. We introduce well-estimated prototypes by including category-wise cross-domain information to link the two alignments: the pixel-level alignment is achieved using the jointly trained style transfer module with the **prototypical semantic consistency**, while the feature-level alignment is enforced to cross-domain features with the **pixel-to-prototype contrast**. Our extensive experiments demonstrate that our method outperforms existing state-of-the-art methods using DeepLab.*

1. Introduction

Semantic segmentation is a fundamental task in computer vision that consists in predicting the class label of each pixel in an image [12]. Segmentation has been the focus of extensive research in the supervised regime, leading to considerable progress in recent years [2, 5, 6, 53]. Much of this progress can be attributed to the availability of large-scale annotated datasets, such as Cityscapes [10] and ADE20K [60]. However, the cost of manual annotation often compels the practitioners to rely on pre-trained models in test environments, without fine-tuning. Unfortunately, these pre-trained models generally perform poorly on test samples that differ from the training data, due to the so-called *domain shift* problem [47]. To address this problem, Domain Adaptive Semantic Segmentation (DASS) methods [11] have been proposed that enable learning on the domain of interest, without needing annotations.

Traditionally, the DASS methods have been designed to address the problem primarily from one of the two fronts: feature-level alignment [20, 23] or pixel-level alignment [8, 35], both aiming to align the labelled source and unlabelled target domain. Very recently, self-training [1, 21] with student-teacher framework [45] has emerged as an effective technique to iteratively fine-tune on the target domain by using the most confident pseudo-labels. With so many genres of existing methods for DASS, it begs the question: ***How to combine the best of the worlds in DASS?***

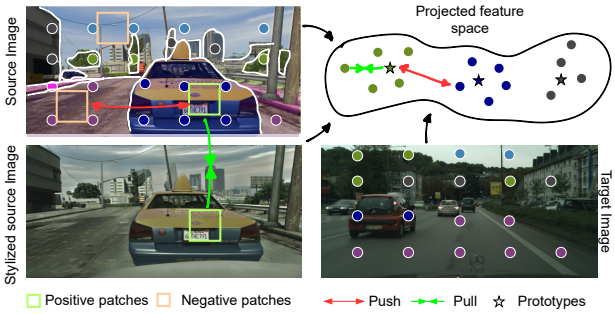

Figure 1. They key idea of our proposed CONFETI is to use contrastive learning to unify feature-level alignment with pixel-level alignment. Features of the pixels from the same class, but across domains, are pulled towards it's corresponding prototype and pushed apart from dissimilar ones. For improved style transfer it enforces that the positive patches in source and stylized images are closer than the negative patches in the projected feature space.

To find answer to this question we turn our attention to the contrastive formulation, InfoNCE [15], that has been found to be effective for a myriad of tasks such as supervised segmentation [51], weakly supervised segmentation [13] and unpaired image translation [37], among others. Given the versatility of the contrastive loss in addressing representation learning and unpaired image translation, both of which have proven to be useful for DASS [18], in this work we propose an unsupervised contrastive learning framework for DASS. We leverage contrastive learning to conduct both *feature-level* and *pixel-level* alignment, while synergistically using the mean-teacher framework.

From the perspective of feature-level alignment, the contrastive loss ensures that the representation of pixels belonging to the same class, but *across* domains, are closer to each other in an embedding space (*i.e.*, intra-class compactness) while being discriminative to other unrelated classes (*i.e.*, inter-class dispersion). Such a formulation comes with two key advantages: (**i**) it enables us to contrast with pixel locations not only from the same image but from other images (both source and target domain); and (**ii**) it allows to consider the global structure present in a scene, which is in sharp contrast to methods (for *e.g.*, self-training) that treat each pixel individually. To reduce computation, we maintain classwise *prototypes* computed from Class Activation Maps [59] (see Sec. 3.2.1 for details), and enforce *pixel-to-prototype* contrast where the pixel embeddings are contrasted with the prototypes instead of pixels.

On the other front of pixel-level alignment, which essentially consists in generating target-*like* source images, the contrastive learning helps in the style transfer by making unpaired image translation one-sided [37], instead of the classical bi-directional cycle-consistent translation [61]. Concretely, we adopt CUT [37] that uses a patchwise contrastive loss to ensure that the feature representation of corresponding patches in the source and target-*like* (or stylized) source image are closer in the embedding space than other random patches. To further improve the stylization, we propose to use a semantic consistency loss that makes sure that the semantic content is not altered during the stylization process (see Sec. 3.2.2 for details).

We call our framework **CON**trastive **FE**a**T**ure and p**I**xel alignment (CONFETI) as it allows to amalgamate both feature-level and pixel-level alignment using the unique formulation of contrastive loss (see Fig. 1). We also show that CONFETI can be seamlessly integrated with the mean-teacher framework, where the prototypes are computed using the teacher network, and the student network learns to match the representation of the corresponding prototype.

In summary, our **contributions** are three-fold: (**i**) We propose an unsupervised contrastive learning framework called CONFETI that enables both feature-level and pixel-level alignment for addressing DASS; (**ii**) We show that CONFETI can easily be integrated with the very effective self-training strategy; and (**iii**) We extensively evaluate our method on standard DASS benchmarks and set new state-of-the-art results when compared with methods that use the common DeepLab [5] segmentation network.

2. Related Works

Domain Alignment in DASS. Following the success of domain *alignment* in image classification, the semantic segmentation methods have adopted various alignment techniques, which ensure that the source and target distributions are aligned at different levels of the pipeline under some metric. Particular to DASS, the three levels are namely latent feature space, input (or pixel) space and output space. First, the feature-level alignment DASS methods seek to minimize the distance between the marginal feature distributions of the source and the target, by either minimizing Maximum Mean Discrepancy along with aligning the correlation matrices [4], or by using a domain discriminator to increase *domain confusion* in the learned features [19, 23, 31, 44, 50]. Second, the pixel-level alignment consists in bridging the domain gap via style transfer [24, 61], which involves transferring the 'appearance' of the target domain onto the source images. The DASS methods that incorporate pixel-level alignment [8,9,35,38,54,55] have proven to be very effective since the content do not change drastically in the DASS benchmarks. Third, the output-level alignment methods circumvent the high dimensionality of the latent feature space and instead perform adversarial adaptation in the output space of the network [36, 49, 50]. The complementary nature of the alignment techniques has led to the development of DASS methods [18, 29, 46] that combine different domain alignments, to better mitigate the domain shift. Our proposed CONFETI also harmoniously combines feature-level alignment with pixel-level alignment, but via contrastive learning [15].

Self-training in DASS. Drawing inspirations from the semi-supervised learning, the idea of using pseudo-labels generated for unlabelled target data, and using them to iteratively fine-tune (or self-train) the target model is also prevalent in DASS [57, 58, 62]. Several of the very recent DASS methods using the self-training strategy have adopted the popular idea of model *ensembling* for obtaining pseudo-labels. In particular, these methods [1, 9, 21] use the mean-teacher [45] (or student teacher) framework where the teacher network, which is an exponential moving average of the student network weights, provides pseudo-labels to train on the target data. In our work we also utilize the teacher network to obtain pseudo-labels, which are then used by the student network for the feature-level alignment.

Contrastive learning in DASS. Learning discriminative visual features in a self-supervised manner, where given an anchor data point, the network must distinguish a similar sample from other dissimilar samples, forms the core idea of contrastive learning [7, 15, 16]. Due to its effectiveness, DASS methods [25, 33, 34, 52] have adopted it for learning compact latent embedding space. For instance, CLST [33] leverages self-training to obtain pseudo-labels for computing class-specific prototypes (or centroids), which are then used in a contrastive manner to learn more compact features. Similarly, ProCA [25] contrasts the pixel representation with the prototypes, except the source prototypes are updated with the target in a moving average fashion. SePiCo [52] goes a step further and estimates the distribution of each prototypes, rather than point estimates. Differ-

ent from the previous works that exploit contrastive learning, our CONFETI computes the prototypes from mixed images, obtained with ClassMix [48], and the class activation maps [59]. Besides feature-level alignment, we also employ contrastive learning for the pixel-level alignment.

3. Methods

In this work we propose **CON**trastive **FE**a**T**ure and p**I**xel alignment (CONFETI), a domain adaptive semantic segmentation (DASS) method, that leverages the contrastive formulation to (**i**) learn a well structured pixel embedding space for feature-level alignment; and (**ii**) foster more accurate style transfer between the source and the target for pixel-level alignment. Before we introduce our method, we formalize the problem and discuss the preliminaries.

3.1. Preliminaries

Problem Definition. We define the input space of images as $\mathcal{X}$, where each image $X \in \mathcal{X}$ is denoted as $X \in \mathbb{R}^{H \times W \times 3}$, H and W being the height and width. The output label space $\mathcal{Y}$ is formed by labels belonging to K semantic categories, such that the segmentation map can be denoted as $Y \in \mathbb{R}^{H \times W}$. In DASS, we are given a source domain dataset with annotations $\mathcal{D}_{\mathrm{S}} = \{(X_i^{\mathrm{S}}, Y_i^{\mathrm{S}})\}_{i=1}^{n_{\mathrm{S}}}$ and an unlabelled target domain dataset $\mathcal{D}_{\mathrm{T}} = \{X_i^{\mathrm{T}}\}_{i=1}^{n_{\mathrm{T}}}$, such that $p(\mathcal{X}^{\mathrm{S}}) \neq p(\mathcal{X}^{\mathrm{T}})$. The objective of DASS is to learn a mapping function $f\colon \mathcal{X} \rightarrow \mathcal{Y}$ that can correctly predict unlabelled target samples. The function $f = f_c \circ f_b$ is modeled by a neural network, such that it is a composition of the backbone feature extractor f_b and the segmentation decoder f_c, which is parameterized by $\theta = \{\theta_b, \theta_c\}$.

Self-training. It has been shown in the DASS literature [21, 48, 52] that self-training (ST) is an effective technique to reduce the domain gap, which we adopt as a baseline. In details, the ST approach uses the student-teacher (or mean teacher [45]) model, where the teacher network $\tilde{f}$ provides pseudo-labels $\hat{Y}$ on-the-fly for the unlabelled target samples to train the student network f:

$$\hat{Y}^{\mathrm{T}}(j) = \arg\max_{c \in \mathcal{Y}} \tilde{p}_j^c \tag{1}$$

where $\tilde{p}_j^c = \tilde{f}(X_j^{\mathrm{T}})$ is the target network prediction probability at pixel j for class c. The pseudo-labelled target data is then used to train the student network using a standard cross-entropy (CE) loss:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{HW}\frac{1}{K}\sum_{j=1}^{H \times W}\sum_{c=1}^{K} \hat{Y}^{\mathrm{T}}(j) \log p_j^c \tag{2}$$

where $p_j^c = f(X_j^{\mathrm{T}})$ is the student network prediction probability for the j^{th} pixel to be belonging to class c and $\hat{Y}^{\mathrm{T}}(j)$ is the corresponding pseudo-label obtained using Eq. (1). Besides the ST objective on the target data, we also optimize the Eq. (2) for the annotated source data using the ground-truth labels Y^{S}.

In ST, the parameters of the teacher network $\tilde{\theta}$ are obtained by taking an exponential moving average (EMA) of the student network parameters θ at every iteration t as:

$$\tilde{\theta}_{t+1} \leftarrow \beta\tilde{\theta}_t + (1-\beta)\theta_t \tag{3}$$

where β is a momentum update hyperparameter, which is in general set to 0.999. Note that we optimize the Eq. (2) on mixed target images that are obtained using the ClassMix [48] augmentation, instead of the real target images, in order to avoid ST with noisy pseudo-labels.

3.2. Contrastive Learning Framework for DASS

In this work we propose CONFETI, a contrastive learning framework that enables both *feature-level* and *pixel-level* alignment using the contrastive formulation InfoNCE [15]. Our choice of using the contrastive formulation is motivated by the fact that InfoNCE has proven to be beneficial for learning compact pixel embedding space for supervised segmentation [51] and accurate style transfer [37]. We argue that compact pixel representation and accurate style transfer are two key ingredients to attain feature-level and pixel-level alignment, respectively. Given, the feature-level and pixel-level alignment are proven to be two essential ingredients for an effective DASS method (*e.g.*, CyCADA [18]), we revisit the two learning fronts by employing contrastive learning and bring them into an unique framework. Below we elaborate in detail the two alignment techniques.

3.2.1 Contrastive feature-level alignment

The feature-level alignment is carried out in the latent feature space of the network by adopting the *pixel-to-prototype* contrast, with the aim of enforcing the pixels of the same semantic category across domains to be close in the embedding space. The driving force behind adopting such a formulation is that the ST training objective only takes into account the 'local' context around a given pixel and completely ignores the 'global' context from other samples in the dataset and beyond. Moreover, being in the unsupervised scenario, *pixel-to-pixel* contrast [51] with noisy pseudo-labels of Eq. (1) will lead to reduced performance.

Guided by these insights we first construct semantic prototypes (or class centroids), one per class, and use these prototypes to pull pixels of the positive class together and at the same time push away pixels from negative classes. Constructing and updating the prototypes is a design choice in itself, and are mainly updated using arithmetic mean [25, 52] or the exponential moving average [25, 56]. However, the direct update of the prototype using all the features is

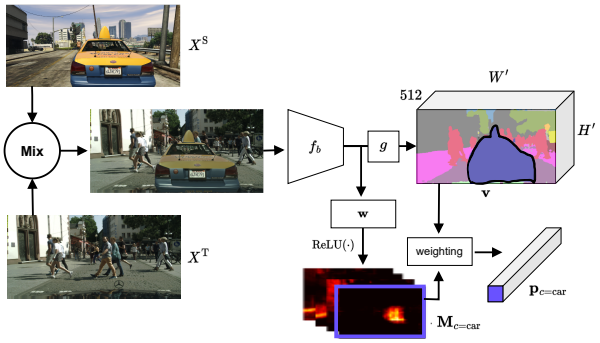

Figure 2. Overview of weighted prototype estimation. The features $\mathbf{v}$ of the mixed image at each spatial location is weighted with the CAM $\mathbf{M}_c$ of each class to obtain a prototype $\mathbf{p}_c$. $\mathbf{v}$ is overlayed with the segmentation map for ease of visualization

unlikely to be applied to the target domain due to the possible erroneous and noisy pseudo-labels, which may lead to inaccurate prototype estimation. Therefore, we employ a weighted prototype estimation method based on the class activation map (CAM) [13, 59].

Weighted prototype estimation. The CAM highlights the most discriminative pixel locations in an image that a network looks at for predicting a given class. We employ the CAM in DASS in order to estimate the prototypes. The idea is to compute a weighted average of the embeddings from pixel locations that are maximally activated by CAM for a given class. This results into a prototype that most likely represents the class under consideration. Although, CAM may fail to highlight precise boundary regions of objects, it does not impact our algorithm as the boundary pixel features do not define the canonical representation of objects.

Concretely, features $\mathbf{f} = f_b(X) \in \mathbb{R}^{H' \times W' \times D}$ are obtained from the feature extractor, followed by applying Global Average Pooling (GAP) to collapse the spatial dimensions; where H', W' and D represent the dimensions of the intermediate feature maps. To get the CAM for a particular class c, a fully connected layer, having parameters $\mathbf{w} \in \mathbb{R}^{K \times D}$, is learned that outputs a score for each class c as:

$$s_c = \frac{1}{H'W'} \sum_{d=1}^{D} \mathbf{w}_{c,d} \sum_{j=1}^{H' \times W'} \mathbf{f}_{d,j} \tag{4}$$

The CAM, denoted as $\mathbf{M}_c$, is then computed for each class as:

$$\mathbf{M}_c = \text{ReLU}\big(\sum_{d=1}^{D} \mathbf{w}_{c,d} \mathbf{f}_{d,:}\big) \tag{5}$$

where the ReLU$(\cdot)$ is applied to ignore all negative values. Note that one CAM $\mathbf{M}_c$ is obtained per image.

As shown in Fig. 2, the prototypes for each class c are then estimated using the just computed $\mathbf{M}_c$ and the projected intermediate features of the images in a mini-batch as:

$$\mathbf{p}'_c = \frac{\sum_{j \in \mathcal{N}_c} \mathbf{M}_{c,j} \mathbf{v}_j}{\sum_{j' \in \mathcal{N}_c} \mathbf{M}_{c,j'}} \tag{6}$$

where $\mathcal{N}_c$ denotes the pixel locations in the entire dataset that correspond to the top-n highest CAM activation values for class c, and $\mathbf{v}_j = g(\mathbf{f}_j) \in \mathbb{R}^{512}$ are the projected features obtained using a non-linear projection head $g(\cdot)$. Note that the prototypes are computed using the teacher network.

The prototypes are then updated in an online manner using the EMA of the prototypes from each mini-batch and the past ones as:

$$\mathbf{p}_c \leftarrow \gamma \mathbf{p}_c + (1 - \gamma) \mathbf{p}'_c \tag{7}$$

where $\mathbf{p}_c$ is the CAM-based prototype of class c for the whole dataset and γ being the momentum update hyperparameter. Importantly, instead of estimating prototypes only on the source domain, we apply the update on the mixed image of the source and target domain. Such an update setting can facilitate the reduction of the domain gap by applying the subsequent contrastive loss, which sets us apart from other works using prototypes [13, 25, 33].

Prototypical contrastive alignment. In order to bring closer the representation of the pixels that belong to the same semantic category, we adopt the prototypical contrastive loss (PCL) [27], originally designed for image-level representation learning. In details, the PCL ensures that the representation of a sample (*pixel* in our case) is more similar to its corresponding prototype than other unrelated ones. Given, our prototypes are computed using the mixed images, they can be seen as a 'bridge' between the source and target domain. Thus, minimizing the PCL translates to aligning the two domains into a shared embedding space. Our proposed method differs from the previous DASS works [25, 52], which also adopt the PCL, in the manner in which the prototypes are computed.

Given a projected feature $\mathbf{v}_j$ extracted by the student $g \circ f_b$ corresponding to a pixel X_j and the estimated prototypes $\mathcal{P} = \{\mathbf{p}_c\}_{c=1}^{K}$, the pixel-to-prototype likelihood for pixel j is given as:

$$p_{j,c} = \frac{\exp(\mathbf{v}_j \cdot \mathbf{p}_c / T)}{\sum_{k=1}^{K} \exp(\mathbf{v}_j \cdot \mathbf{p}_k / T)} \tag{8}$$

where $\mathbf{p}_c$ is the prototype belonging to the same class as pixel X_j and T is the temperature. To attract the feature to prototype of class c and repel it from others, the following loss is adopted:

$$\mathcal{L}_{\text{PCL}} = -\frac{1}{N} \sum_{j=1}^{N} Y(j, c) \log p_{j,c} \tag{9}$$

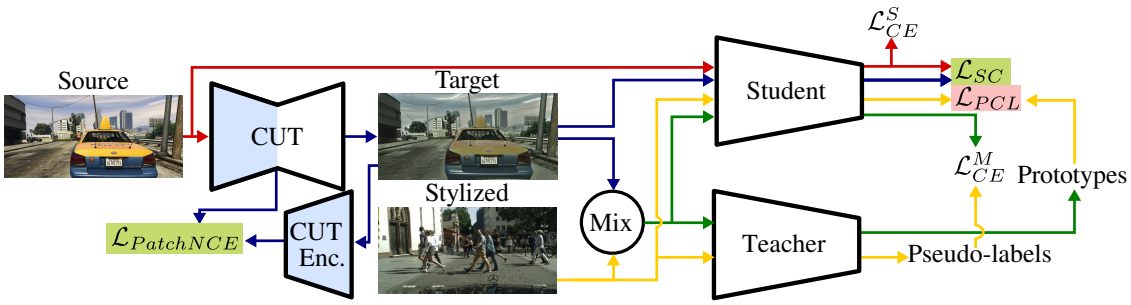

Figure 3. Overview of the proposed **CONFETI**, which is a mean-teacher framework that unifies feature alignment with pixel alignment. The feature alignment exploits the prototypical contrastive loss ($\mathcal{L}_{\text{PCL}}$) with mixed-prototypes to align the domains in the feature space. The pixel-alignment exploits contrastive learning ($\mathcal{L}_{\text{PatchNCE}}$ and $\mathcal{L}_{\text{SC}}$) to carry out style transfer, aligning the two domains in the pixel space

where N is the number of randomly sampled features to avoid performance degradation caused by erroneous labels, and Y is the ground truth mask for the source domain features and pseudo-labels for the target domain features.

3.2.2 Contrastive pixel-level alignment

In order to further mitigate the domain-shift, CONFETI enables pixel-level alignment by generating target-*like* source samples that have the same content as the source images but appear to be drawn from the target domain. As shown in the previous generative DASS approaches [18, 42, 43], pixel-level alignment can sometimes outperform feature-level alignment methods.

A commonality among these generative DASS approaches is the use of cycle-consistency loss [61] that is susceptible to two major issues: (**i**) the generator encodes noise (or high frequency signal) during the forward translation, which is then utilized as a *shortcut* during the reverse translation to reconstruct the original image, and (**ii**) the generator can accurately translate images which adhere to the target domain statistics but dramatically changing the source content. To overcome the drawbacks of bi-directional translation, one-sided unpaired image translation have been proposed [3, 14, 37, 41]. In this work we adopt CUT [37], an unpaired image translation (or *stylization)* method that exploits the contrastive learning to associate corresponding *patches* in the two domains to be similar.

Concretely, as shown in Fig. 1 we employ the patch-based InfoNCE loss [37], where given the projected features of an anchor patch in the stylized image $\hat{\mathbf{z}}^{\text{S}\to\text{T}}$, the corresponding *positive* patch in the source image $\mathbf{z}^{\text{S}}_{+}$, and a set of *negative* patches from other locations in the source image $\{\mathbf{z}^{\text{S}}_{-}\}_{j=1}^{n}$ (see the Supplement for details), is given as:

$$\mathcal{L}_{\text{PatchNCE}} = -\mathbb{E}_{\hat{\mathbf{z}}^{\text{S}\to\text{T}}} \log \frac{\exp(\hat{\mathbf{z}}^{\text{S}\to\text{T}} \cdot \mathbf{z}^{\text{S}}_{+}/T)}{\left[\exp(\hat{\mathbf{z}}^{\text{S}\to\text{T}} \cdot \mathbf{z}^{\text{S}}_{+}/T) + \sum_{j=1}^{n} \exp(\hat{\mathbf{z}}^{\text{S}\to\text{T}} \cdot \mathbf{z}^{\text{S}}_{-,j}/T)\right]} \tag{10}$$

To further alleviate the problem of semantic inconsistency in the translated images we propose a joint training of the stylization and the segmentation module. It follows a *virtuous cycle*: better segmentation model leads to high quality image-translation, and better quality domain translation leads to improved segmentation under domain-shift. In details, we additionally propose to use a prototypical semantic consistency loss that uses the prototypes (detailed in Sec. 3.2.1) to ensure that the target-*like* images do not hallucinate incorrect content during translation, which are not present in the source:

$$\mathcal{L}_{\text{SC}} = \frac{1}{H'W'K} \sum_{j=1}^{H'\times W'} \sum_{c=1}^{K} \|\mathbf{v}^{\text{S}}_{j} \cdot \mathbf{p}_c - \phi(\hat{\mathbf{v}}^{\text{S}\to\text{T}}_{j}) \cdot \mathbf{p}_c\|^2 \tag{11}$$

where $\phi(\cdot)$ is a learnable affine transformation applied on the features from the stylized image to allow gaps between two images, for instance color, lightness and texture.

We further use the prototypes in Eq. (11), which models each category, to classify the features extracted by the student's backbone. The semantic consistency loss thus ensures that each pixel corresponds to an identical class after the translation. Note that the gradients from this loss are not used for the optimization of the segmentation network.

3.2.3 Training objectives

The whole pipeline of CONFETI is depicted in the Fig. 3. The training will be divided into two phases. In the first, or the *joint* training phase, the overall loss $\mathcal{L}_{\text{Joint}}$ is given as:

$$\mathcal{L}_{\text{Joint}} = \underbrace{\mathcal{L}^{\text{S,M}}_{\text{CE}}}_{\text{self-training}} + \underbrace{\lambda_{\text{PCL}}\mathcal{L}^{\text{S,M}}_{\text{PCL}}}_{\text{feature alignment}} + \underbrace{\lambda_{\text{style}}(\mathcal{L}_{\text{PatchNCE}} + \mathcal{L}_{\text{SC}})}_{\text{pixel alignment}} \tag{12}$$

where S and M denote that the losses are applied to the source and the mixed images, respectively. λ_{PCL} and λ_{style} are used for weighing the corresponding losses.

In order to avoid overfitting the segmentation model to the stylized image features, such as textures, during the joint

Table 1. Comparison results on **GTA5 → Cityscapes**. Methods based on pixel alignment are highlighted with colors. † indicate methods trained at higher resolution. The best performance are in **bold** and the best performance in low resolution setting are marked with underline

Methods	road	sideway	building	wall	fence	pole	light	sign	vegetation	terrace	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
Cycada [18]	86.7	35.6	80.1	19.8	17.5	38.0	39.9	41.5	82.7	27.9	73.6	64.9	19.0	65.0	12.0	28.6	4.5	31.1	42.0	42.7
Li *et al.* [29]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
FDA-MBT [55]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
DACS [48]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
CPSL [28]	91.7	52.9	83.6	43.0	32.3	43.7	51.3	42.8	85.4	37.6	81.1	69.5	30.0	88.1	44.1	59.9	24.9	47.2	48.4	55.7
Ma *et al.* [32]	92.5	58.3	86.5	27.4	28.8	38.1	46.7	42.5	85.4	38.4	91.8	66.4	37.0	87.8	40.7	52.4	**44.6**	41.7	59.0	56.1
ProCA [25]	91.9	48.4	87.3	41.5	31.8	41.9	47.9	36.7	86.5	42.3	84.7	68.4	43.1	88.1	39.6	48.8	40.6	43.6	56.9	56.3
DecoupleNet [26]	88.5	47.8	87.4	38.3	36.9	44.9	53.8	39.6	88.0	38.7	88.8	70.4	39.4	87.8	31.4	55.0	37.4	47.1	55.9	56.7
ProDA [56]	87.8	56.0	79.7	**46.3**	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
SePiCo [52]	95.2	67.8	88.7	41.4	38.4	43.4	55.5	63.2	88.6	46.4	88.3	73.1	49.0	91.4	**63.2**	60.4	0.0	45.2	60.0	61.0
CONFETI (Ours)	95.7	69.9	89.5	34.6	42.6	40.9	57.5	59.4	88.6	**49.0**	88.2	72.8	**53.4**	90.1	61.8	54.9	13.9	**50.2**	63.4	62.2
HRDA [22]†	96.2	73.1	**89.7**	43.2	39.9	47.5	60.0	60.0	**89.9**	47.1	**90.2**	75.9	49.0	91.8	61.9	59.3	10.2	47.0	**65.3**	63.0
CONFETI (Ours)†	**96.5**	**75.6**	88.9	45.1	**45.9**	**50.1**	**61.2**	**68.2**	89.4	45.7	86.3	**76.3**	49.9	**92.2**	55.1	**62.8**	16.7	33.8	63.1	**63.3**

training with the image-to-image translation model, we propose a second-round training where we train the segmentor from scratch, with the style transfer network kept frozen.

4. Experiments

4.1. Implementation Details

Datasets. We follow the experimental protocols adopted in the previous DASS works [25, 48, 52, 56]. For the source domain, we use the synthetic GTA dataset [39] containing 24,966 synthetic images of resolution 1914×1052 and the SYNTHIA dataset [40] with 9400 synthetic images of resolution 1280×760. As target domain we use the Cityscapes dataset [10] which contains 2975 training and 500 test images of resolution 2048×1024. In the low-resolution setting, following [52], images are resized to 1280×640 for Cityscapes dataset and to 1280×720 for GTA dataset before randomly cropping to 640×640. For fair comparison with HRDA [22], we also perform experiments in full resolution and use 1024×1024 crops for training.

Training. As in [25, 48, 52, 56], we adopt DeepLab-V2 [5] with ResNet-101 [17] as backbone. We use the AdamW optimizer [30] with the initial learning rate set to 6×10^{-5} and weight decay of 0.01. We adopt the warm-up policy as well as the rare class sampling proposed by [21]. Following [48], we apply the color jittering, Gaussian blurring and ClassMix [48] on the mixed images.

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

We compare with recent state-of-the-art methods [18, 22, 25, 26, 28, 29, 32, 48, 52, 55, 56], especially those using style transfer [18, 29, 32, 55] and prototypes [25, 52, 56]. On the GTA → Cityscapes task, CONFETI is compared separately with HRDA [22] since it operates at full resolution.

The quantitative comparison on GTA → Cityscapes is reported in Tab. 1. We observe that our approach outperforms prior methods with the mIoU of 62.2%. In particular, our method shows its high capacity on easy to confuse class pairs, such as motor-bike, road-sideway, and person-rider pairs. Moreover, working at higher resolution improves the performance on small objects, especially on challenging classes such as 'train', with an overall improvement of 0.3% over HRDA [22]. As shown in the Tab. 2 for the SYNTHIA → Cityscapes benchmark, the proposed method also obtains state-of-the-art performance. More precisely, we obtain 58.7% and 67.4% mIoU in the 16-category and 13-category evaluation protocols, respectively. In both the benchmarks, we observe that existing methods based on style transfer underperform more recent methods based on feature alignment and self-training. Overall our CONFETI demonstrates that unifying these two orthogonal research directions can lead to state-of-the-art results.

4.3. Ablation Studies

Evaluation of the proposed pipeline. We thoroughly ablate our proposed CONFETI in order to measure the impact of: (**i**) joint training of the style-transfer and segmentation models, (**ii**) our two-stage training procedure, and (**iii**) the introduction of prototypes in our pixel-alignment technique. In these ablations, we start from the self-training baseline which is described in Sec. 3.1.

We report the results of the ablations in Tab. 3. First, when the style transfer network, which aligns domains at pixel-level, is trained separately from the segmentation network (see model **A** in Tab. 3), we observe a performance drop of 0.4% mIoU w.r.t the baseline. It indicates that stylization is not able to bridge the domain gap. On the contrary, when we perform feature alignment via prototypical contrastive learning but without any stylization, we observe a clear gain of +2.5% (see model **B**). Surprisingly, including

Table 2. Comparison results on **SYNTHIA** → **Cityscapes**. Methods based on pixel alignment are highlighted with colors

	road	sideway	building	wall*	fence*	pole*	light	sign	vegetation	sky	person	rider	car	bus	motor	bike	mIoU	mIoU*
Li *et al.* [29]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
FDA-MBT [55]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	38.4	51.1	-	52.5
DACS [48]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	**90.8**	67.6	38.3	83.0	38.9	28.5	47.6	48.3	54.8
Ma *et al.* [32]	75.7	30.0	81.9	11.5	2.5	35.3	18.0	32.7	86.2	90.1	65.1	33.2	83.3	36.5	35.3	54.3	48.2	55.5
ProCA [25]	**90.5**	**52.1**	84.6	29.2	3.3	40.3	37.4	27.3	86.4	85.9	69.8	28.7	88.7	53.7	14.8	54.8	53.0	59.6
CPSL [28]	87.3	44.4	83.8	25.0	0.4	42.9	47.5	32.4	86.5	83.3	69.6	29.1	89.4	52.1	42.6	54.1	54.4	61.7
ProDA [56]	87.8	45.7	84.6	**37.1**	0.6	44.0	54.6	37.0	**88.1**	84.4	74.2	24.3	88.2	51.1	40.5	45.6	55.5	62.0
DecoupleNet [26]	77.8	48.6	75.6	32.0	1.9	**44.4**	52.9	38.5	87.8	88.1	71.1	34.3	88.7	58.8	50.2	61.4	57.0	64.1
SePiCo [52]	77.0	35.3	85.1	23.9	3.4	38.0	51.0	55.1	85.6	80.5	73.5	46.3	87.6	**69.7**	**50.9**	**66.5**	58.1	66.5
CONFETI (Ours)	83.8	44.6	**86.9**	15.4	**3.7**	44.3	**56.9**	**55.5**	84.9	86.2	**73.8**	**46.8**	**90.1**	57.1	46.0	63.2	**58.7**	**67.4**

Table 3. Ablation study on the **GTA5** → **Cityscapes**. PCL denotes the use of prototypical contrastive learning

Method	Style Transfer: Offline	Joint	Two-stage	PCL	mIoU	Δ
Baseline					57.5	-
A	✓				57.1	-0.4
B				✓	60.0	+2.5
C	✓			✓	57.6	+0.1
D		✓			59.0	+1.5
E		✓	✓		59.0	+1.5
CONFETI (Ours)		✓	✓	✓	**62.2**	**+4.7**

Table 4. Ablation on the **GTA** → **Cityscapes** benchmark. **Top**: Comparison with alternative style transfer methods. **Bottom**: Impact of the various semantic consistency losses

	FDA [55]	GPA [32]	AdaIN [24]	Ours
mIoU	60.5	59.4	59.1	**62.2**
	w/o $\mathcal{L}_{SC}$	$\mathcal{L}_{CE}$	MSE	$\mathcal{L}_{SC}$ (Ours)
mIoU	58.0	58.6	61.6	**62.2**

offline stylization into the pipeline (see model **C**) is again detrimental (57.6% vs +60.0%). This may be explained by inaccurate stylization if not jointly done, which alters the image content and makes the estimated prototypes noisy.

When the stylization is no longer trained offline but jointly with the segmentation model (see model **D**), we start observing gains in performance (+1.5% compared to the baseline). Then, our two-stage training without $\mathcal{L}_{PCL}$, *i.e.*, model **E**, does not improve the performance w.r.t model **D**, which shows the limits of stand-alone pixel-level alignment. However, when we combine joint style transfer and prototypical contrastive in a two stage training fashion, CONFETI achieves the best performance of 62.2%, which is +4.7% higher than the baseline. These ablations justify that both the feature-level and pixel-level alignment contribute constructively by reducing the domain gap: joint training of the style transfer model improves pixel alignment and prototype estimation while the feature alignment loss promotes domain invariant features.

Comparison with style transfer methods. We now extend our comparison by evaluating alternative style-transfer methods to gauge the advantage of contrastive style transfer in CONFETI. We consider the neural network-based AdaIN [24] and training-free methods (*e.g.*, FDA [55], GPA [32]). In these experiments, we replace the CUT module in our CONFETI with each alternative method and employ the prototypical contrastive loss $\mathcal{L}_{PCL}$ in Eq. (9) as the training objective. We report the quantitative results in the top half of Tab. 4 and the qualitative results in Fig. 4. We can observe that our CONFETI empirically outperforms all the aforementioned style transfer competitors. From this qualitative comparison, we observe that FDA or GPA generally generate image with poor quality and many artifacts. On the contrary, our method outputs images where the content of the source image is preserved but the style is well-transferred. The difference is especially clear in the sky region where the most methods fail.

Evaluation of the semantic consistency loss. We compare our proposed semantic consistency loss (*i.e.*, $\mathcal{L}_{SC}$ in Eq. (11)) for pixel-level alignment with some other alternative losses from the DASS literature. First, we consider a baseline without consistency loss (referred to as *w/o* $\mathcal{L}_{SC}$). Then, we consider a baseline, referred to as $\mathcal{L}_{CE}$, which is inspired by CyCADA [18] and uses the cross-entropy between the estimated segmentations. Finally, we include a variant of CONFETI where $\mathcal{L}_{SC}$ is replaced by a MSE loss. The quantitative results in the bottom half of Tab. 4 show that the last two variants that operate at the feature level clearly perform better. Among the two variants of CONFETI, the contrastive formulation attains higher performance. The qualitative results are shown in the Fig. 4.

Effect of mixed prototype estimation. We now analyse the design choices for prototype estimation. The prototypes

Figure 4. Qualitative comparison of different style-transfer methods and semantic consistency losses in the **GTA** → **Cityscapes** setting

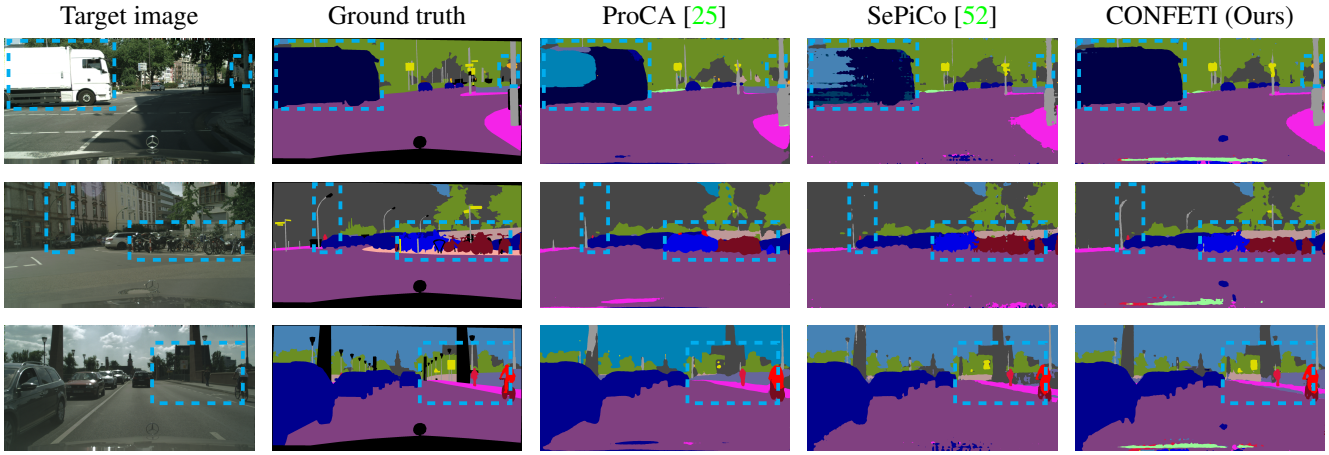

Figure 5. Qualitative comparison of the segmentation maps of CONFETI with the state-of-the-art methods [25,52] on **GTA** → **Cityscapes**

Table 5. Impact of design choices for prototypes estimation in terms of mIoU on the **GTA** → **Cityscapes** benchmark

Method	w/o CUT	CUT	w/o CAM	w/ CAM
w/o prototype	57.5	57.1	-	-
Source prototype	58.6	60.5	-	-
Mixed prototype	60.0	62.2	60.1	62.2

can be estimated from the original source images or from the mixed images obtained by ClassMix [48]. These two solutions are compared to a baseline where the prototypes are not used. We perform experiments with and without style transfer with CUT and report the results in the left of Tab. 5. First, we observe that estimating the prototypes with features from the mixed images leads to higher mIoUs. This result shows that including cross-domain information in the prototypes helps adaptation. Furthermore, when the mixed prototypes are used, combing with CUT further boosts the performance (62.2% vs 60.0%) showing that the prototypes with rich cross-domain information improve both pixel- and feature-level alignment.

Effect of CAM-based weighting. We now validate our CAM-based solution to estimate the weighted prototypes (described in Sec. 3.2.1). In the right of Tab. 5, we compare the performance of models where the prototypes are estimated with and without the proposed CAM-based weighting technique and trained with CUT. These results demonstrate the effectiveness of the weighted prototypes as they outperform the unweighted prototypes by $+1.1\%$ of mIoU.

5. Conclusion

In this work, we presented CONFETI, a novel approach that unifies feature-level and pixel-level cross-domain alignment. CONFETI was integrated with the mean-teacher self-training framework and was shown to improve the performance of DASS through the joint use of prototypical contrastive loss and style transfer. On one hand, our jointly trained style transfer module produced high-quality stylized images that bridge the domain gap and aid in prototype estimation. On the other hand, the well-estimated prototypes, in conjunction with the prototypical contrastive loss, further reinforce the feature-level alignment and enhance the DASS performance. Both quantitative and qualitative experiments demonstrated the effectiveness of CONFETI, outperforming existing state-of-the-art methods.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 1, 2

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 1

[3] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in Neural Information Processing Systems*, 30, 2017. 5

[4] Róger Bermúdez-Chacón, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. A domain-adaptive two-stream u-net for electron microscopy image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 400–404. IEEE, 2018. 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*, 40(4):834–848, 2017. 1, 2, 6

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[8] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019. 1, 2

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019. 2

[10] 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*, pages 3213–3223, 2016. 1, 6

[11] Gabriela Csurka, Riccardo Volpi, and Boris Chidlovskii. Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey. *arXiv preprint arXiv:2112.03241*, 2021. 1

[12] Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 14(1-2):1–162, 2022. 1

[13] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 1, 4

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 5

[15] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 1, 2, 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[17] 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*, pages 770–778, 2016. 6

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998. Pmlr, 2018. 1, 2, 3, 5, 6, 7

[19] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[20] 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*, pages 1335–1344, 2018. 1

[21] 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*, pages 9924–9935, 2022. 1, 2, 3, 6

[22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 6

[23] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 590–605, 2018. 1, 2

[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 7, 8

[25] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 36–54. Springer, 2022. 2, 3, 4, 6, 7, 8

[26] Xin Lai, Zhuotao Tian, Xiaogang Xu, Yingcong Chen, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Decouplenet: Decoupled network for domain adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 369–387. Springer, 2022. 6, 7

[27] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[28] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11593–11603, 2022. 6, 7

[29] 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*, pages 6936–6945, 2019. 2, 6, 7

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

[31] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2

[32] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2021. 6, 7, 8

[33] Robert A Marsden, Alexander Bartler, Mario Döbler, and Bin Yang. Contrastive learning and self-training for unsupervised domain adaptation in semantic segmentation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 2, 4

[34] 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*, pages 12435–12445, 2021. 2

[35] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 1, 2

[36] 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*, pages 3764–3773, 2020. 2

[37] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 1, 2, 3, 5, 8

[38] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2990–2998, 2020. 2

[39] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 6

[40] 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*, pages 3234–3243, 2016. 6

[41] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Nicu Sebe, and Elisa Ricci. Trigan: Image-to-image translation for multi-source domain adaptation. *Machine Vision and Applications*, 32:1–12, 2021. 5

[42] Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8099–8108, 2018. 5

[43] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 5

[44] Tong Shen, Dong Gong, Wei Zhang, Chunhua Shen, and Tao Mei. Regularizing proxies with multi-adversarial training for unsupervised domain-adaptive semantic segmentation. *arXiv preprint arXiv:1907.12282*, 2019. 2

[45] 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*, 30, 2017. 1, 2, 3

[46] Marco Toldo, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for mobile semantic segmentation based on cycle consistency and feature alignment. *Image and Vision Computing*, 95:103889, 2020. 2

[47] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011. 1

[48] 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*, pages 1379–1389, 2021. 3, 6, 7, 8

[49] 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*, pages 1456–1465, 2019. 2

[50] 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*, pages 2517–2526, 2019. 2

[51] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 1, 3

[52] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 4, 6, 7, 8

[53] 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*, 34:12077–12090, 2021. 1

[54] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020. 2

[55] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 6, 7, 8

[56] 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*, pages 12414–12424, 2021. 3, 6, 7

[57] Yangsong Zhang, Subhankar Roy, Hongtao Lu, Elisa Ricci, and Stéphane Lathuilière. Cooperative self-training for multi-target adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5604–5613, 2023. 2

[58] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021. 2

[59] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 3, 4

[60] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 1

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017. 2, 5

[62] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 2