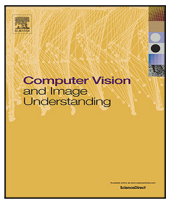
Computer Vision and Image Understanding 260 (2025) 104464

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Context perturbation: A Consistent alignment approach for Domain Adaptive Semantic Segmentation

Meiqin Liu [a], Zilin Wang [a], Chao Yao [b,*], Yao Zhao [a], Wei Wang [a], Yunchao Wei [a]

[a] School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China
[b] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

ARTICLE INFO

Communicated by Yu-Chiang Frank Wang

Keywords:
Domain adaptation
Semantic segmentation
Consistency regularization
Representation learning
Self-training

ABSTRACT

Domain Adaptive Semantic Segmentation (DASS) aims to adapt a pre-trained segmentation model from a labeled source domain to an unlabeled target domain. Previous approaches usually address the domain gap by consistency regularization which is implemented based on the augmented data. However, as the augmentations are often performed at the input level with simple linear transformations, the feature representations suffer limited perturbation from these augmented views. As a result, they are not effective for cross-domain consistency learning. In this work, we propose a new augmentation method, namely contextual augmentation, and combine it with contrastive learning approaches from both the pixel and class levels to achieve consistency regularization. We term this methodology as Context Perturbation for DASS (CoPDASeg). Specifically, contextual augmentation first combines domain information by class mix and then randomly crops two patches with an overlapping region. To achieve consistency regularization with the two augmented patches, we focus on both pixel and class perspectives and propose two parallel contrastive learning paradigms (*i.e.*, pixel-level contrastive learning and class-level contrastive learning). The former aligns the pixel-to-pixel feature representations, and later aligns class prototypes across domains. Experimental results on representative benchmarks (*i.e.*, **GTA5 → Cityscapes** and **SYNTHIA → Cityscapes**) demonstrate that CoPDASeg improves the segmentation performance over state-of-the-arts by a large margin.

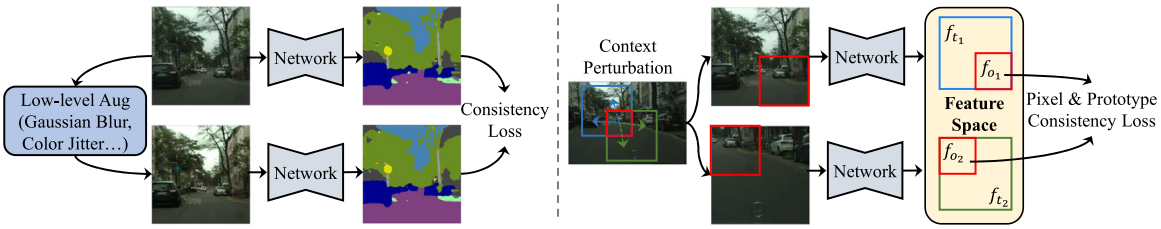
1. Introduction

Semantic segmentation (Long et al., 2015; Chen et al., 2017; Zhao et al., 2017; Wang et al., 2019; Yuan et al., 2019; Chen and Hu, 2021; Xu et al., 2023) is one of the most critical tasks in computer vision. It aims to predict the per-pixel classification of an input image. Thanks to the advancements in deep learning (Krizhevsky et al., 2017; Long et al., 2015; Huang et al., 2017; Song et al., 2019), the performance of semantic segmentation has witnessed remarkable improvements (Chen et al., 2017; Huang et al., 2019; Liu et al., 2019) in recent years. However, the state-of-the-art methods require large amounts of pixel-wise annotations from diverse scenarios to enhance their generalization capabilities. The process of pixel-wise annotation is labor-intensive and time-consuming. For example, annotating a single high-resolution image from the Cityscapes (Cordts et al., 2016) dataset consumes more than 1.5 h on average. One promising direction to circumvent this problem is to use more accessible synthetic datasets, such as GTA5 (Richter et al., 2016) and SYNTHIA (Ros et al., 2016). Unfortunately, models trained on synthetic data face challenges when adapted to real data on account of the domain gap. The domain gap is caused by various factors like layout, appearance, *etc*. Domain Adaptive Semantic Segmentation is considered as a plausible solution to ease the synthetic-to-real domain gap, hence reducing the annotation cost.

The initial researches in DASS focus on addressing the domain gap through domain adversarial training. Specifically, these methods aim to align the distribution between source and target domains at different levels, such as feature-level (Chen et al., 2019b; Pan et al., 2020; Kim and Byun, 2020), output-level (Melas-Kyriazi and Manrai, 2021; Luo et al., 2019; Tsai et al., 2018; Chen et al., 2021) and image-level (Dundar et al., 2020; Hoffman et al., 2018; Yang and Soatto, 2020). However, these methods strive for marginal distribution alignment between the source and target images at the cost of ignoring the crucial aspect of semantic consistency. As a consequence, it potentially leads to semantic confusion between different classes, particularly in cases where target supervision is absent. Furthermore, adversarial training often introduces the use of an additional domain discriminator or style-transfer network, which introduces additional computational costs.

* Corresponding author.
E-mail address: yaochao@ustb.edu.cn (C. Yao).

https://doi.org/10.1016/j.cviu.2025.104464
Received 16 September 2024; Received in revised form 27 May 2025; Accepted 13 August 2025
Available online 25 August 2025



(a) Consistency Training with Low-Level Augmentation (b) Consistency Training with Context Perturbation

Fig. 1. The objective of consistency training is to guarantee the consistent predictions of the model regardless of the input perturbations, thereby enhancing the robustness of the learned model. (a) Previous methods for consistency training usually employ low-level augmentations (flipping, color jittering, Gaussian blurring, *etc*) as perturbations for unlabeled images. (b) Our proposed context perturbation uses different contexts as the perturbations.

Subsequent researches employ the self-training strategy (Tranheden et al., 2021; Lian et al., 2019; Zhang et al., 2017; Zou et al., 2019) to minimize semantic confusion. The pseudo labels are produced according to various clues, *e.g.*, classifier confidence (Corbiere et al., 2021; Zou et al., 2019; Wang et al., 2021b), feature distance (Zhang et al., 2021), *etc*. More recent studies employ consistency regularization (Tranheden et al., 2021; Araslanov and Roth, 2021) to mitigate the bias in pseudo labeling. It reinforces the consistency between two views, namely the original view and the corresponding augmented view, as shown in Fig. 1(a). These augmentations primarily encompass linear transformations, such as kernel filters, color space transformation, geometric transformation and so on. These manually designed methods are effective, reproducible and reliable for encoding color and geometric space invariance within the original dataset. However, recent researches on self-supervised learning (Zhang et al., 2019; Gidaris et al., 2018) have uncovered that these low-level transformations pose little change and are effortlessly accommodated (*i.e.*, overfitted) by deep neural networks. This observation highlights the potential insufficiency of basic image processing methods in effectively perturbing the input distribution.

To solve this issue, in this paper, we devise more effective perturbations and propose contextual augmentation to strengthen consistency regularization, as shown in Fig. 1(b). Specifically, contextual augmentation mixes the images from two domains and crops two patches with a certain overlapping region. The underlying rationale is that the learned representations are influenced by varying contexts, even within the same region. Contextual augmentation introduces more robust perturbations in the feature space compared to basic image manipulation, thereby providing enhanced benefits for subsequent consistency training. To achieve consistency regularization, we consider it from two perspectives and realize them with two contrastive learning paradigms, *i.e.*, pixel-to-pixel and prototype-to-prototype contrastive learning. Pixel-to-pixel contrastive learning aligns pixel features at corresponding locations within the overlapped regions. Prototype-to-prototype contrastive learning facilitates the alignment of prototypes belonging to the same class across different domains. As a result, these two contrastive learning schemes work together to promote consistency under contextual augmentation, effectively mitigating the domain gap. The proposed method, namely Context Perturbation for DASS (CoPDASeg), adopts a one-stage pipeline without extra special training techniques, which is simple but effective. Different from Lai et al. (2021), Chen et al. (2023) achieving it on the unlabeled domain, our contextual augmentation is based on the mixed labeled and unlabeled domain. The feature space is modeled on the mixed domain, which is facilitated to achieve cross-domain alignment in subsequent pixel-level alignment. In addition, we propose the prototype-to-prototype contrastive loss to further enhance cross-domain semantic consistency.

Extensive experiments and ablation studies are conducted on representative benchmarks for DASS, *i.e.*, **GTA5 → Cityscapes** and **SYNTHIA → Cityscapes**. The results show that our approach consistently outperforms the state-of-the-art methods. To be specific, on hard classes with lower frequency, our method boosts the performance by a large margin. The main contributions of this paper can be summarized as:

- We propose a new augmentation method, *i.e.*, contextual augmentation, which introduces strong and effective perturbations for cross-domain consistency regularization.
- To achieve consistency regularization under contextual augmentation, we propose pixel-to-pixel and prototype-to-prototype contrastive losses, which align feature representation at both the pixel and class levels.
- Extensive experiments on popular semantic segmentation benchmarks show that the proposed CoPDASeg achieves superior performance on the target domain over state-of-the-arts. It especially shows outstanding results on long-tailed classes such as “motorbike”, “train”, “light”, *etc*.

2. Related work

2.1. Semantic segmentation

Semantic segmentation is a critical task in computer vision that aims to divide an image into non-overlapping regions representing various semantic categories. The advent of deep learning leads to remarkable advances in semantic segmentation, with Long et al. (2015) pioneering the use of convolutional neural networks for this task. Some methods further improve performance by enlarging receptive fields or capturing context information. For example, Chen et al. (2017) propose dilated convolution to enlarge the receptive field of filters. It incorporates a larger context without increasing the number of parameters or the amount of computation. Zhao et al. (2017) propose a pyramid pooling module that fuses different scale features to capture context information. Wang et al. (2019) propose to use CNNs for low-level feature extraction and a Structured Random Forest (SRF)-based border ownership detector for high-level feature extraction. In recent years, attention-based Transformer (Vaswani et al., 2017) has emerged as a highly successful approach in the field of Natural Language Processing (NLP). Taking inspiration from this, Transformers have been widely adopted in various visual tasks, such as image classification (Liu et al., 2021; Dosovitskiy et al., 2021), object detection (Carion et al., 2020) and semantic segmentation (Cheng et al., 2021, 2022). Unfortunately, although Transformers show superior performance compared to ResNet backbones, their effectiveness heavily relies on a substantial amount of pixel-wise annotations. For example, Cityscapes (Cordts et al., 2016) and PASCAL VOC (Everingham et al., 2010) consist of more than 5K and 13K annotated images respectively. To solve this issue, some synthetic datasets are proposed such as GTA5 (Richter et al., 2016) and SYNTHIA (Ros et al., 2016). However, there is a domain gap between real-world datasets and synthetic datasets. As a result, models trained solely on synthetic datasets tend to exhibit low performance when applied to real-world datasets. In this work, we aim to learn an adaptive model that aligns the distribution between source and target domains with only the source domain supervision.

2.2. Domain adaptive semantic segmentation

Currently, the DASS methods can be mainly categorized into two groups: the adversarial training ones and the self-training ones. Specifically, the adversarial training network contains two parts which are the generator and discriminator. The generator is responsible for generating the dense predictions or acting as the feature extractor. Meanwhile, the discriminator aims to distinguish the domain of the features (Chen et al., 2019b; Pan et al., 2020; Kim and Byun, 2020), final outputs (Melas-Kyriazi and Manrai, 2021; Luo et al., 2019; Tsai et al., 2018), or images (Dundar et al., 2020; Hoffman et al., 2018; Yang and Soatto, 2020). In the AdaptSeg framework (Tsai et al., 2018), adversarial training is implemented in the multi-level output space. Tsai et al. (2018) align the segmentation results of the source and target domains by the adversarial network. Luo et al. (2019) incorporate class-wise information into their methodology and introduce a category-level adversarial network. This approach reduces the weight of the adversarial loss for category-level aligned features while simultaneously increasing the adversarial force for poorly aligned ones. Chen et al. (2021) propose a classification constrained discriminator, which not only addresses the adversarial training issue but also mitigates the problem of feature distortion. However, most previous methods usually focus on intra-class distributional alignment. By contrast, we attempt to explore the relationship between clusters of different categories. We set a generic semantic-guided prototype contrast learning method to enhance class-wise discriminative information. It minimizes intra-class discrepancy and maximizes inter-class margin across the two domains.

Self-training methods typically involve two components: a teacher network and a student network. The teacher network is initially trained on a labeled source dataset and subsequently utilized to generate high-confidence pseudo labels for the unlabeled target domain. These pseudo labels are then employed for training the student network. However, it is noted that pseudo labels frequently contain noise, resulting in reduced reliability in terms of their confidence level. Therefore, many previous studies (Jiang et al., 2022) strive to enhance the quality of pseudo labels and minimize the impact of noise. Mei et al. (2020) propose a dynamic approach to estimate a threshold for each semantic category of pseudo labels with the intent of decreasing the percentage of hard classes. Zhang et al. (2021) calculate the feature distance between a sample point and all prototypes, and then reweight its corresponding pseudo logits accordingly to effectively eliminate noise. Wang et al. (2021a) leverage the guidance from self-supervised depth estimation, which is available on both domains, to diminish the domain gap. However, the majority of existing methods include complex multi-stage processes and rely on various training techniques. In contrast, our framework accomplishes one-stage and end-to-end adaptation without the need for separate pre-processing stages or training techniques.

2.3. Consistency regularization

The key idea behind consistency regularization is that predictions on unlabeled instances should not change significantly to perturbations. In semi-supervised learning, perturbations are usually derived from image augmentations (Xie et al., 2020; Sohn et al., 2020), *e.g.* kernel filters, color space transformation, geometric transformation, *etc*. For DASS, some methods (Chen et al., 2019a; Zhou et al., 2022) utilize consistency regularizer to minimize distribution discrepancies at image-level. DACS (Tranheden et al., 2021) enforces consistency between predictions of target and mixed domains (*i.e.*, mixing source and target domain images by class-mix Olsson et al., 2021). SAC (Araslanov and Roth, 2021) employs standard low-level data augmentations, including photometric noise, flipping, scaling, *etc*, and ensures consistency of the semantic predictions across these image transformations. However, low-level image augmentation is unable to generate significant perturbations in the feature space. As a consequence, this limits the performance gain brought by consistency regularization. To solve this issue, we propose contextual augmentation method which introduces stronger perturbations in the feature space. Additionally, cross-domain consistency regularization is employed to boost the performance of DASS. Lai et al. (2021), Chen et al. (2023) propose to utilize different contexts to perturb the feature space only on the unlabeled data. Different from (Lai et al., 2021; Chen et al., 2023), our contextual augmentation is based on the cross-domain mixed image. It aims to achieve cross-domain alignment by subsequently pixel-level alignment. In addition, we propose the prototype-to-prototype contrastive loss to further enhance semantic consistency between domains.

2.4. Contrastive learning

Contrastive learning strategy (Oord et al., 2018) has gained attention for its promising ability to learn representations without explicit supervision. By pulling positive pairs closer and pushing negative pairs apart, it has the capacity to greatly enhance representation learning. In the semi-supervised semantic segmentation task, Alonso et al. (2021) apply pixel-level contrastive training to yield similar feature representations for intra-class samples across the whole dataset. For DASS, Jiang et al. (2022) construct intra-class and inter-class relations by applying the contrastive loss on the different class prototypes. Different from previous methods (Jiang et al., 2022; Alonso et al., 2021), we construct pixel-level and class-level contrastive paradigms. It enhances both the pixel-wise and class-wise consistency of the overlapping regions in different contexts, respectively.

3. Method

We first introduce our pipeline in Section 3.1. Next, in Section 3.2, we describe the process of our proposed contextual augmentation. Finally, in Section 3.3 and Section 3.4, we present our pixel-to-pixel and prototype-to-prototype contrastive learning strategies, respectively.

3.1. Overview

In the setting of DASS, we have a labeled source domain $D_S=(x_s, y_s)$ and an unlabeled target domain $D_T=(x_t)$, where x_s and x_t are the input images and y_s is the corresponding annotation of source image. Both source domain D_S and target domain D_T share the same C classes. There are two kinds of images in the input batch, *i.e.*, labeled source images x_s and unlabeled target images x_t.

To facilitate the adaptation of the network, our proposed domain adaptation pipeline is based on a self-training network (Tarvainen and Valpola, 2017) which consists of two different branches, namely the student network Θ and the teacher network Θ'. The student network is composed of a feature extractor Θ_E and a classifier Θ_C. In addition, an additional projection head Θ_P is introduced and constitutes the auxiliary networks $\Theta_P(\Theta_E)$ to map the features to the latent space.

The labeled source images x_s are randomly cropped and fed into the student network Θ to get the predictions $p_s = \Theta_C(\Theta_E(x_s))$, which is supervised by the ground truth y_s with the cross-entropy (CE) loss:

$$\mathcal{L}_s = -\sum_{i=1}^{H\times W}\sum_{c=1}^{C} y_s^{(i,c)} \log \Theta(x_s)^{(i,c)} \tag{1}$$

where H and W represent the height and the width of an image, C denotes the number of classes, i and c indicate the ith pixel and cth class in image and classes, respectively.

On the other hand, to enhance the domain adaptation, pseudo labels $\hat{y}_T$ are generated for target samples x_t with the teacher network Θ'. Additionally, a series of data augmentations are applied to the target domain images x_t and obtain new augmented images, denoted as x_t^A. Similarly, the pseudo labels $\hat{y}_t$ are transformed into augmented pseudo labels $\hat{y}_t{}^A$ at the same time. The augmented samples x_t^A along with their corresponding augmented pseudo labels $\hat{y}_t^A$ are used as training samples

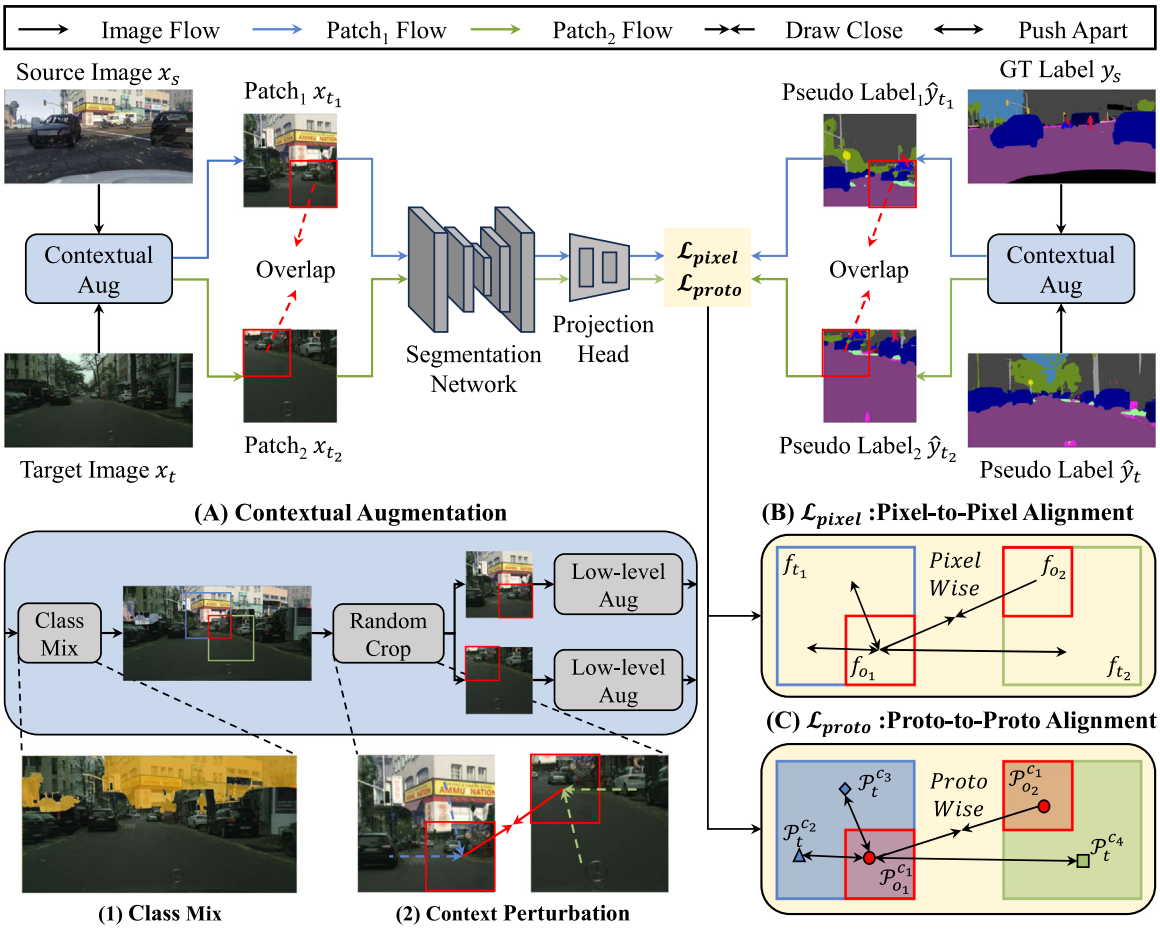


Fig. 2. Overview of CoPDASeg. It includes the source domain branch and the target domain branch. The target images x_t and source images x_s pass through contextual augmentation **(A)** to get two patches x_{t_1} and x_{t_2} with an overlapping region x_o. We achieve consistent alignment on x_o by our proposed two contrastive losses, *i.e.*, $\mathcal{L}_{pixel}$ **(B)** and $\mathcal{L}_{proto}$ **(C)**. In **(A)**, following DACS (Tranheden et al., 2021), a mixed image x_m is firstly created by having one set of pixels coming from x_s, and one set of pixels coming from x_t. Then, we randomly crop two patches with an overlapping region from x_m. Our purpose is to use different context information to perturb the overlapping region feature *i.e.*, context perturbation. In **(B)**, we regularize the pixel-wise consistency between f_{o_1} and f_{o_2} by computing pixel-to-pixel contrast: impelling positive-pair embeddings closer, *i.e.*, two features f_{o_1} and f_{o_2} of the same overlapping region x_o derived from F_{pos_1} and F_{pos_2}, and pushing away the negative embeddings, *i.e.*, the corresponding negative set F_{neg}. In **(C)**, we use two features f_{o_1} and f_{o_2} to yield the corresponding prototypes $\mathcal{P}^c_{o_1}$ and $\mathcal{P}^c_{o_2}$ to provide semantic guidance. And then, we regularize the prototype-wise consistency between $\mathcal{P}^c_{o_1}$ and $\mathcal{P}^c_{o_2}$. Similarly, we apply prototype-to-prototype contrastive learning, which draws positive pair close, *i.e.*, two prototypes with the same class closer, while pushing negative pairs apart, *i.e.*, prototype $\mathcal{P}^c_{o_1}$ and $\mathcal{P}^c_{o_2}$ keep away from global prototype $\mathcal{P}^c$ with different classes.

for Θ. The self-training loss, implemented with the cross-entropy loss, is then computed as shown below:

$$\mathcal{L}_t = -\sum_{i=1}^{H\times W}\sum_{c=1}^{C} q_t(\hat{y}_t^A)^{(i,c)} \log \Theta(x_t^A)^{(i,c)} \tag{2}$$

where q_{t_1} denotes the ratio of pseudo label pixels exceeding a threshold τ of the maximum softmax probability. Following the approach of DACS (Tranheden et al., 2021), to ensure the learning of domain-robust features, color jitter, Gaussian blur and ClassMix are adopted as data augmentations. The auxiliary networks $\Theta_P(\Theta_E)$ are trained on the two patches x_{t_1} and x_{t_2} by enforcing a consistency of features *i.e.*, f_{t_1} and f_{t_2}. This objective is achieved by minimizing our proposed pixel-to-pixel and prototype-to-prototype contrastive loss $\mathcal{L}_{pixel}$ and $\mathcal{L}_{proto}$. More details are presented in Sections 3.3 and 3.4.

As to the teacher network, it is noted that no gradients are back-propagated into the teacher network. Thus, following the previous method (Tarvainen and Valpola, 2017), the weights of teacher network Θ' are set as the Exponential Moving Average (EMA) of the weights of the student network at each iteration t.

$$\Theta' \leftarrow \alpha\Theta' + (1-\alpha)\Theta \tag{3}$$

where α is a momentum parameter.

3.2. Contextual augmentation

The objective of consistency regularization is to enforce an invariance of the model's predictions to various perturbations, facilitating the learning of robust representations. Instead of simple linear transformation (Araslanov and Roth, 2021) on original images, we propose context perturbations to more effectively disrupt the features and then promote model robustness by the subsequent consistency regularization.

Specifically, we devise a novel data augmentation, namely contextual augmentation. Given a source image x_s and a target image x_t, a binary mask M is first generated by randomly selecting half of the classes in the source image x_s. This mask M is utilized to mix the source and the target images to generate a mixed image x_m, which is formulated as:

$$x_m = M \cdot x_s + (1-M)\cdot x_t \tag{4}$$

Subsequently, we crop two random patches x_{m_1} and x_{m_2} from mixed images x_m which are confined to have an overlapping region x_o. Finally, x_{m_1} and x_{m_2} are augmented to get x_{t_1} and x_{t_2} by a series of image

augmentations, which include random flip, color jitter and Gaussian blur. The procedure for contextual augmentation is shown in Fig. 2(A). Similarly, the annotations for the patches $\hat{y}_{t_1}$ and $\hat{y}_{t_2}$ are also augmented to be in line with augmented input images.

Our contextual augmentation aims to help the network produce more robust features against varying environments. As a consequence, it effectively reduces the influence of noise existing in the pseudo label during the process of self-training and results in considerably better performance, as shown in Section 4.2. And each component of our contextual augmentation is discussed in Table 4.

3.3. Pixel-to-pixel consistent alignment

To achieve consistency alignment with contextual augmentation, we take inspiration from contrastive learning and propose pixel-to-pixel contrastive loss. It creates contrastive pairs based on the inter-pixel positional information and achieves pixel-to-pixel feature alignment across domains.

Specifically, as shown in Fig. 2, given augmented samples x_{t_1} and x_{t_2} created by contextual augmentation, latent features f_{t_1} and f_{t_2} are extracted from x_{t_1} and x_{t_2} by feature extractor Θ and projection head Θ_P. Where the features of the overlapping region x_o in f_{t_1} and f_{t_2} are denoted as f_{o_1} and f_{o_2}, respectively. The projection head here serves as an information bottleneck to preserve the informative contextual features. In Table 3, the experiments are conducted to highlight the contribution of the projection head.

To realize the pixel-to-pixel contrastive learning, two features $f^i_{o_1}$ and $f^i_{o_2}$ at the same location of f_{o_1} and f_{o_2} are set as a positive pair, as shown in Fig. 3. In addition, the corresponding negative set F^i_n is also constructed for the ith feature $f^i_{o_1}$. Therefore for the ith feature $f^i_{o_1}$, the contrastive loss is be defined as:

$$\mathcal{L}_c(f^i_{o_1}, f^i_{o_2}) = \log \frac{\gamma_1(f^i_{o_1} \cdot f^i_{o_2})}{\gamma_1(f^i_{o_1} \cdot f^i_{o_2}) + \sum_{f_n \in F^i_n} \gamma_1(f^i_{o_1} \cdot f_n)} \tag{5}$$

where f_n is the negative pair of $f^i_{o_1}$ and it is sampled from F^i_n, γ_1 denotes the exponential function of the cosine similarity c between two features with a scaling factor s_1:

$$\gamma_1(f^i_{o_1} \cdot f^i_{o_2}) = exp(s_1 \cdot c(f^i_{o_1}, f^i_{o_2})) \tag{6}$$

The construction of a negative sample set F^i_n is crucial for contrastive learning. The most straightforward method is to take all the other pixels as negative samples. However, for the dense predictions on all the pixels, there are many pixels from different positions belonging to the same class even the same object. It may easily lead to false negative pairs between the same class pixels, especially the background class or large objects such as the sky and sidewalk. To avoid this, we sample all the feature vectors whose classes are different from the $f^i_{o_1}$ as a negative sample set F^i_n. All features come from each mini-batch during training:

$$F^i_n = \{f_n | \hat{y}_n \neq \hat{y}^i_{o_1}\} \tag{7}$$

where $\hat{y}^i_{o_1}$ and $\hat{y}_n$ denote the pseudo label of $f^i_{o_1}$ and f_n. In this way, the features from the different contexts of two patches are leveraged to enhance consistency between the overlapping region features $f^i_{o_1}$ and $f^i_{o_2}$. As a result, it helps to establish a discriminative inter-class feature representation in feature space, which is especially important for DASS. Empirically, we observe that more negative samples lead to better performance for contrastive learning. Thus, the negative set F_n is expanded. It is not only from the current image but also from all the unlabeled images within the current training batch. Moreover, a memory bank is maintained to further increase negative samples. It stores the features in the past few batches to get sufficient negative samples.

However, the pseudo label produced for the unlabeled target image is very noisy, the low-quality noisy samples may have a bad influence on model optimization. If all the positive feature pairs in the overlapping region are used to calculate the $\mathcal{L}_c$, the less confident pairs f_{o_1} and f_{o_2} may corrupt the model. Therefore, to avoid this, a positive set F_{p_1} is also constructed for f_{o_1} to filter out those less confident positive samples from f_{o_2}. Specifically, the maximum probability is computed among all the classes in f_{o_1} and f_{o_2}, *i.e.*, $p_{o_1} = max(\Theta_C(f_{o_1}))$ and $p_{o_2} = max(\Theta_C(f_{o_2}))$, as the confidence of each feature f_{o_1} and f_{o_2}. And a threshold τ is set. Only the features f_{o_2} whose confidence is higher than τ are saved as positive samples. To prevent the more confident features from corrupting towards the less confident ones, we further filter out the less confident feature f_{o_2} than f_{o_1} from the positive set F_{p_1}. Therefore, the finally positive set F_{p_1} which are filtered out from f_{o_1} are defined as:

$$F_{p_1} = \{f^i_{o_1} | (p^i_{o_1} < p^i_{o_2}) \wedge (p^i_{o_2} > \tau)\} \tag{8}$$

Similarly, for the features of the second patch f_{t_2}, the positive set F_{p_2} are obtained in the same way. We minimize the $\mathcal{L}_c$ on the feature from F_{p_1} and F_{p_2}. The pixel-to-pixel contrastive loss $\mathcal{L}_{pixel}$ is defined as:

$$\mathcal{L}_{pixel} = -\frac{1}{N_1} \sum_{f^i_{o_1} \in F_{p_1}} \mathcal{L}_c(f^i_{o_1}, f^i_{o_2}) - \frac{1}{N_2} \sum_{f^i_{o_2} \in F_{p_2}} \mathcal{L}_c(f^i_{o_2}, f^i_{o_1}) \tag{9}$$

where N_1 and N_2 denote the number of elements in the set F_{p_1} and F_{p_2}, respectively.

In summary, the pixel-to-pixel contrastive loss (Eq. (9)) ensures that pixel-level features at the same position are aligned, while features at different positions are separated. It progressively enhances the consistency of pixel-level features under different contexts during training. Consequently, pixel-to-pixel contrastive learning increases the segmentation model robustness against varying environments.

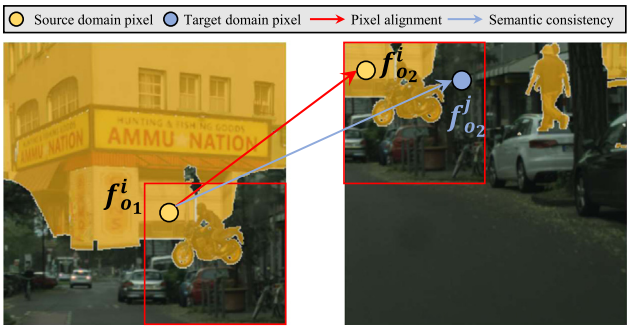


Fig. 3. Visualization of the process of pixel-to-pixel alignment on the two augmented patches derived from our contextual augmentation method.

3.4. Prototype-to-prototype consistent alignment

The proposed pixel-to-pixel contrastive loss (Eq. (9)) focuses on the consistency between corresponding positive pixel pairs of f_{o_1} and f_{o_2} in the overlapping region. We further consider cross-domain semantic consistency for pixels of the same class but from different domains on the mixed image (*e.g.*, $f^i_{o_1}$ and $f^j_{o_2}$), as illustrated in Fig. 3. To accomplish semantic consistency, we propose a prototype-to-prototype contrastive loss that aligns the prototypes belonging to the same class. It utilizes the semantic concept as a guide to enforce feature representation alignment across domains.

Specifically, the overlapping feature f_{o_1} and f_{o_2} are used to construct two class-dependent feature prototypes $\mathcal{P}^c_{o_1}$ and $\mathcal{P}^c_{o_2}$ as category centroids. A prototype for the class c is assembled as an average of selected features whose confidence exceeds the threshold τ:

$$\mathcal{P}^c_{o_1} = \frac{\sum_{i=1}^{N} f^i_{o_1} \cdot \mathbb{1}[p^i_c > \tau]}{\sum_{i=1}^{N} \mathbb{1}[p^i_c > \tau]} \tag{10}$$

Algorithm 1 CoPDASeg Algorithm.

Input: Source-domain and target-domain images x_s and x_t, student network Θ, teacher network Θ', maximum/warm-up iteration L/L_w and hyperparameters λ_{consis}, λ_{FD}.

Initialize Θ_E with ImageNet pre-trained parameters and randomly initialize two heads Θ_C and Θ_P.
Teacher network init: $\Theta'_E \leftarrow \Theta_E$, $\Theta'_C \leftarrow \Theta_C$.
for $iter \leftarrow 0$ to L **do**
  Randomly sample a source image x_s with y_s and a target image x_t.
  x_t^A, $\hat{y}_t^A \leftarrow$Apply augmentation with $(x_t, \hat{y}_t)$, where $\hat{y}_t$ is generated from $\hat{y}_t \leftarrow \Theta'(x_t)$.
  Compute predictions $\hat{Y}_s \leftarrow \Theta'(x_s)$, $\hat{Y}_m \leftarrow \Theta'(x_m)$.
  Train Θ, using Eq. (1), Eq. (2).
  if $iter > L_w$ **then**
    x_{t_1}, x_{t_2}, $\hat{y}_{t_1}$, $\hat{y}_{t_2} \leftarrow$Apply contextual augmentation with $(x_t, \hat{y}_t)$
    Compute hidden-layer feature maps f_{t_1} and f_{t_2}.
    Train Θ, Θ_P via Eq. (9), Eq. (13).
  end if
  Update Θ' with Θ via Eq. (3).
end for
return Final segmentation network Θ.

where N denotes the number of pixels from features f_{o_1}, $\mathbb{1}[\cdot]$ is an indicator function, which equals to 1 if the predicted probability p_c^i is greater than τ and 0 otherwise. Similarly, $\mathcal{P}_{o_2}^c$ is constructed in the same way. We enforce the consistency between $\mathcal{P}_{o_1}^c$ and $\mathcal{P}_{o_2}^c$. In addition, to more accurately depict the category centroids, the global prototype $\mathcal{P}_t$ is calculated from all the unlabeled images within the current training batch:

$$\mathcal{P}_t^c = \frac{\sum_{i=1}^{N} f_t^i \cdot \mathbb{1}[p_c^i > \tau]}{\sum_{i=1}^{N} \mathbb{1}[p_c^i > \tau]} \tag{11}$$

where f_t^i includes the features f_{t_1} and f_{t_2}. The global prototype $\mathcal{P}_t$ is used as negative pairs when we enforce the consistency between the local prototype of the same class. Specifically, for the prototype $\mathcal{P}_{o_1}$, the same class prototype $\mathcal{P}_{o_1}^c$ and $\mathcal{P}_{o_2}^c$ are positive pairs and all the other class global prototypes $\mathcal{P}_t^k$ are the negative pairs:

$$\mathcal{L}_c(\mathcal{P}_{o_1}^c, \mathcal{P}_{o_2}^c) = log \frac{\gamma_2(\mathcal{P}_{o_1}^c \cdot \mathcal{P}_{o_2}^c)}{\gamma_2(\mathcal{P}_{o_1}^c \cdot \mathcal{P}_{o_2}^c) + \sum_{k \neq c}^{C} \gamma_2(\mathcal{P}_{o_1}^c \cdot \mathcal{P}_t^k)} \tag{12}$$

where γ_2 denotes the exponential function of the cosine similarity between two prototypes with a scaling factor s_2. Therefore the final prototype-to-prototype contrastive loss $\mathcal{L}_{proto}$ is defined as:

$$\mathcal{L}_{proto} = -\sum_{c=1}^{C} (\mathcal{L}_c(\mathcal{P}_{o_1}^c, \mathcal{P}_{o_2}^c) + \mathcal{L}_c(\mathcal{P}_{o_2}^c, \mathcal{P}_{o_1}^c)) \tag{13}$$

The prototype of the augmented image serves as a bridge between the source and target domains, as it shares the same semantic information from both. Specifically, by optimizing the prototype-to-prototype contrastive loss (Eq. (13)), we minimize the intra-category discrepancy and maximize the inter-category margin within the two domains. It facilitates the explicit transfer of knowledge between the source and target domains. The features of the same category in both domains are mixed during prototype-to-prototype contrastive learning, resulting in more accurate class-wise alignments across domains. This, in turn, enables the model to generalize better in both domains.

In conclusion, the procedure of our CoPDASeg is presented in Algorithm 1. A batch of images and labels, x_s, x_t^A and y_s, y_t^A, are used to train the student network Θ by optimizing Eq. (1), Eq. (2). To stabilize the training process, a warm-up iteration step L_w is set. After L_w steps, our proposed two contrastive losses Eqs. (9) and (13) are added into the training process to achieve CoPDASeg. Thus, our final loss function is defined as:

$$\mathcal{L}_{final} = \mathcal{L}_s + \mathcal{L}_t + \lambda_{consis}(\mathcal{L}_{pixel} + \mathcal{L}_{proto}) + \lambda_{FD}\mathcal{L}_{FD} \tag{14}$$

where $\mathcal{L}_s$ denotes the cross-entropy loss on the labeled source domain, $\mathcal{L}_t$ represents self-training loss on the unlabeled target domain, $\mathcal{L}_{FD}$ is thing-class ImageNet feature distance of DAformer (Hoyer et al., 2022a) to further improve results, λ_{consis} and λ_{FD} indicate the hyper-parameters to balance different losses.

4. Experiments

4.1. Experimental setups

Datasets. We evaluate our method on two popular synthetic-to-real benchmarks: **GTA5 → Cityscapes** and **SYNTHIA → Cityscapes**. GTA5 (Richter et al., 2016) is an image dataset synthesized by the physically-based rendered computer game "Grand Theft Auto V". It contains 24,966 city scene images with a resolution of 1914 × 1052 and is shared with the same 19 classes as Cityscapes. SYNTHIA (Ros et al., 2016) is a synthetic urban scene dataset. Following (Tsai et al., 2018), we select its subset, called SYNTHIA-RAND-CITYSCAPES, which has 16 common semantic annotations with Cityscapes. In total, the SYNTHIA dataset contains 9400 images with a resolution of 1280 × 760. GTA5 and SYNTHIA are used as source domain data for training. Cityscapes (Cordts et al., 2016) is a dataset of real urban scenes taken from 50 cities in Germany and neighboring countries, which contains 2975 training and 500 validation urban scene images with a resolution of 2048 × 1024. For the DASS task, we use its training set images as the unlabeled target domain and validation set images for evaluation.

Network Architectures. For the baseline, we select the recent state-of-the-art framework MIC (Hoyer et al., 2023) which is based on DAFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b), which are all based on Transformer architecture. To verify the generalization ability of our method on CNN structures, following previous methods (Wang et al., 2023a,b), we also replace the transformer-based model structure with ResNet101 (He et al., 2016) + DeepLabV2 (Chen et al., 2017). In all architectures, Atrous Spatial Pyramid Pooling (ASPP) with dilated rates {6, 12, 18, 24} (Chen et al., 2017) is used as the segmentation head. An up-sampling layer is used to compute the final per-pixel predictions with the same image size as the input. The projection head (Wang et al., 2021c) is integrated into the network that maps high-dimensional pixel embedding into a 128-d l_2-normalized feature vector. It consists of two 1 × 1 convolutional layers and one intermediate ReLU layer. For a fair comparison, all the backbones are pre-trained on ImageNet (Deng et al., 2009), with the remaining layers being initialized randomly.

Implementation Details. We implement CoPDASeg based on the mmsegmentation toolbox with PyTorch on 4 T V100 GPUs. We apply AdamW (Loshchilov and Hutter, 2019) optimizer with the initial learning rate of $\eta_{base} = 6 \times 10^{-5}$ for the encoder and 6×10^{-4} for the decoder, a weight decay of 0.01, betas (0.9,0.999), linear learning rate warm up with $t_{warm} = 1500$ and linear decay afterward. The input image is resized to 1280 × 720 for GTA and 1280 × 760 for SYNTHIA. For the target domain Cityscapes, the image size is resized to 1280 × 640. We train the network with a batch of four 640 × 640 random crops for a total of 60000 iterations. To stabilize training, our method starts from 5000 iterations, *i.e.*, L_w=5000. Following previous works (Hoyer et al., 2022a,b), for the hyper-parameters in our method, we set λ_{consis}=0.01, λ_{FD}=0.005, momentum β=0.999 and τ=0.95 in all our experiments. In contextual augmentation, the Intersection-over-Union (IoU) range of these two patches is supposed to be within the range [0.1,1.0].

Testing. At the test stage, we only resize the validation images to the same resolution of 1280 × 640 as the input image. Note that there is no extra inference step inserted into the basic segmentation model, that is, the teacher network and projection head are directly discarded.

Table 1
GTA5 (Richter et al., 2016) → Cityscapes (Cordts et al., 2016) adaptation results. We compare our method with state-of-the-art competitors. † denotes using the distillation technique. In all tables, the best result is highlighted in red. The second-best results are highlighted in blue. This table contains three sets of experiments based on DeepLabV2 (Chen et al., 2017), DAFormer (Hoyer et al., 2022a), and HRDA (Hoyer et al., 2022b), respectively. The following tables are also the same (Li et al., 2022b).

Network	Methods	road	side	buil	wall	fence	pole	light	sign	veg	terr	sky	pers	rider	car	truck	bus	train	mbike	bike	mIoU
DeepLabV2	Source Only	70.2	14.6	71.3	24.1	15.3	25.5	32.1	13.5	82.9	25.1	78.0	56.2	33.3	76.3	26.6	29.8	12.3	28.5	18.0	38.6
	AdaptSeg Tsai et al. (2018)	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
	ADVENT Vu et al. (2019)	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
	DACS Tranheden et al. (2021)	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
	EFA Chen et al. (2021)	91.3	54.5	84.9	31.0	25.7	36.3	42.0	33.2	85.0	39.1	86.9	61.2	30.7	83.9	32.6	41.6	5.4	31.1	30.7	48.8
	ProDA† Zhang et al. (2021)	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
	LISR (Li et al., 2022b)	93.7	53.6	83.	5 35.1	21.1	28.6	36.2	42.0	82.2	32.4	86.5	47.3	19.4	83.8	26.0	30.7	30.2	13.1	32.1	46.2
	CPSL† Li et al. (2022a)	92.3	59.9	84.9	45.7	29.7	**52.8**	61.5	59.5	87.9	41.5	85.0	73.0	35.5	90.4	48.7	**73.9**	**26.3**	**53.8**	53.9	60.8
	BLV Wang et al. (2023a)	94.9	68.2	88.8	40.9	37.1	42.6	52.1	62.1	88.3	43.3	89.3	68.6	44.5	88.9	**56.0**	54.6	3.8	38.6	58.3	59.0
	MIC Hoyer et al. (2023)	**96.5**	**74.3**	**90.4**	**47.1**	42.8	50.3	**61.7**	**62.3**	**90.3**	49.2	**90.7**	**77.8**	**53.2**	**93.0**	**66.2**	68.0	6.8	38.0	60.6	**64.2**
	RTea (Zhao et al., 2023)	95.4	67.1	87.9	46.1	44.0	46.0	53.8	59.5	89.7	**49.8**	**89.8**	71.5	40.5	**90.8**	55.0	57.9	22.1	47.7	**62.5**	61.9
	CoPDASeg (Ours)	**95.7**	**71.1**	**90.9**	**52.8**	**50.5**	**53.3**	**66.0**	**65.2**	**90.5**	**52.8**	89.1	**80.0**	**57.4**	90.2	44.7	**72.8**	**44.1**	**60.0**	**65.5**	**68.1**
DAFormer	DAFormer Hoyer et al. (2022a)	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
	BLV Wang et al. (2023a)	96.2	73.1	89.3	53.6	**55.7**	50.9	55.7	61.1	89.7	**52.4**	92.3	74.7	43.5	91.6	74.6	77.4	69.2	58.9	62.3	69.6
	MIC Hoyer et al. (2023)	**96.7**	**75.0**	90.0	**58.2**	50.4	51.1	56.7	62.1	90.2	51.3	92.9	72.4	**47.1**	92.8	**78.9**	83.4	75.6	54.2	62.6	70.6
	PiPa (Chen et al., 2023)	96.1	72.0	90.3	56.6	52.0	55.1	**61.8**	63.7	90.8	**52.6**	**93.6**	74.3	43.6	**93.5**	78.4	**84.2**	**77.3**	**59.9**	**66.7**	71.7
	CDAC (Wang et al., 2023b)	96.5	73.9	89.5	56.8	48.9	50.7	55.8	63.3	89.9	49.1	91.2	72.2	45.4	92.7	78.3	82.9	67.5	55.2	**63.4**	69.6
	RTea (Zhao et al., 2023)	96.1	71.7	89.1	57.8	50.4	**55.9**	59.3	**66.7**	**90.4**	48.2	**94.5**	**74.8**	46.5	**93.8**	78.7	81.6	65.8	57.1	62.8	70.6
	MICDrop (Yang et al., 2024)	**96.5**	74.2	**90.8**	**60.5**	52.0	**55.8**	59.9	65.6	90.3	51.8	93.0	73.1	46.9	93.4	**82.0**	85.8	74.3	56.6	62.8	**71.8**
	CoPDASeg (Ours)	**96.5**	**74.7**	**90.6**	56.4	**52.3**	**55.9**	**61.0**	**68.8**	**91.0**	49.6	91.7	**75.5**	**51.2**	92.7	74.9	**86.9**	**78.2**	**59.7**	**66.0**	**72.3**
HRDA	HRDA Hoyer et al. (2022b)	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
	BLV Wang et al. (2023a)	96.7	76.6	91.5	61.2	56.9	59.4	62.2	**72.8**	91.5	51.2	94.3	77.5	54.7	93.5	83.2	84.7	79.7	**68.1**	67.6	74.9
	MIC Hoyer et al. (2023)	97.4	80.1	91.7	61.2	56.9	59.7	66.0	71.3	91.7	51.4	94.3	79.8	56.1	**94.6**	85.4	90.3	80.4	64.5	**68.5**	75.9
	PiPa (Chen et al., 2023)	96.8	76.3	91.6	63.0	57.7	60.0	65.4	72.6	91.7	51.8	**94.8**	79.7	56.4	94.4	85.9	88.4	78.9	63.5	67.2	75.6
	CDAC (Wang et al., 2023b)	97.1	78.7	**91.8**	59.6	57.1	59.1	66.1	72.2	**91.8**	**53.1**	94.5	79.4	51.6	94.6	84.9	87.8	78.7	64.9	67.6	75.3
	RTea (Zhao et al., 2023)	97.1	75.2	**92.6**	**63.5**	51.8	58.2	**66.5**	71.2	91.1	49.0	**96.8**	**81.5**	54.2	94.2	84.8	86.6	75.7	62.2	66.7	74.7
	MICDrop (Yang et al., 2024)	**97.6**	**81.5**	92.0	62.8	59.4	**62.6**	62.9	**73.6**	91.6	52.6	94.1	80.2	**57.0**	94.8	**87.4**	**90.7**	**81.6**	**65.3**	67.8	**76.6**
	CoPDASeg (Ours)	**97.9**	**82.4**	91.6	**65.0**	**60.8**	**61.1**	**66.7**	67.1	**91.8**	**53.3**	94.1	**81.7**	**59.3**	**95.3**	**88.1**	**91.7**	**83.5**	63.1	**69.1**	**77.1**

Table 2
SYNTHIA (Ros et al., 2016) → Cityscapes (Cordts et al., 2016) adaptation results. The mIoU and the mIoU* indicate mean IoU over 16 and 13 categories, respectively. Category with * stands for three categories not calculated in mIoU*.

Network	Methods	road	side	buil	wall*	fence*	pole*	light	sign	veg	sky	pers	rider	car	bus	mbike	bike	mIoU	mIoU*
DeepLabV2	Source Only	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
	AdaptSeg (Tsai et al., 2018)	79.2	37.2	78.8	10.5	0.3	25.1	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	39.5	45.9
	ADVENT (Vu et al., 2019)	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
	DACS (Tranheden et al., 2021)	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	**90.8**	67.6	38.3	82.9	38.9	28.5	47.6	48.3	54.8
	EFA (Chen et al., 2021)	74.4	28.8	81.5	13.5	1.2	32.6	21.6	32.4	81.5	83.7	52.8	25.8	78.0	30.0	29.6	52.7	45.0	51.8
	ProDA† (Zhang et al., 2021)	**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
	LISR (Li et al., 2022b)	84.6	40.3	74.5	0.5	0.1	27.7	25.4	25.1	78.0	81.8	58.0	19.4	70.5	24.3	17.7	41.5	41.8	49.3
	CPSL† (Li et al., 2022a)	87.2	43.9	85.5	**33.6**	0.3	47.7	57.4	37.2	87.8	88.5	**79.0**	32.0	**90.6**	49.4	50.8	59.8	57.9	65.3
	BLV (Wang et al., 2023a)	70.4	28.9	**89.2**	25.2	**19.9**	40.2	55.2	50.3	86.9	84.2	76.4	40.5	79.6	51.3	49.2	61.2	56.8	63.3
	MIC (Hoyer et al., 2023)	84.7	45.7	88.3	29.9	2.8	**53.3**	**61.0**	**59.5**	86.9	88.8	78.2	**53.3**	89.4	**58.8**	**56.0**	**68.3**	**62.8**	**70.7**
	RTea (Zhao et al., 2023)	**93.2**	**59.6**	86.3	31.3	4.8	43.1	41.8	44.0	**88.6**	**90.5**	70.4	42.6	89.5	56.7	40.2	59.9	58.9	66.4
	CoPDASeg (Ours)	78.6	36.5	**88.6**	28.3	**8.6**	**48.9**	**63.9**	**57.3**	**88.1**	88.8	**79.9**	**59.9**	**92.9**	**82.3**	**61.4**	**66.0**	**64.4**	**72.6**
DAFormer	DAFormer (Hoyer et al., 2022a)	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	89.8	73.2	48.2	87.2	53.2	53.9	61.7	60.9	67.4
	BLV (Wang et al., 2023a)	**86.7**	44.9	89.0	43.2	6.4	52.1	60.0	54.9	**88.2**	91.3	74.9	46.1	88.6	55.6	55.0	62.3	62.5	69.0
	MIC (Hoyer et al., 2023)	83.0	40.9	88.2	37.6	**9.0**	52.4	56.0	56.5	87.6	**93.4**	74.2	**51.4**	87.1	59.6	57.9	61.2	62.2	69.0
	PiPa (Chen et al., 2023)	**87.9**	**48.9**	88.7	45.1	4.5	**53.1**	59.1	**58.8**	87.8	92.2	**75.7**	49.6	**88.8**	53.5	**58.0**	**62.8**	**63.4**	**70.1**
	CDAC (Wang et al., 2023b)	83.7	42.9	87.4	39.8	7.5	50.7	55.7	53.5	85.9	90.9	74.5	47.2	86.0	**60.2**	57.8	60.8	61.5	68.2
	RTea (Zhao et al., 2023)	85.9	43.2	**90.1**	45.1	6.3	52.4	**60.5**	57.1	87.8	**92.2**	75.3	**51.8**	87.4	55.9	54.1	**62.6**	63.0	69.5
	MICDrop (Yang et al., 2024)	81.0	37.1	**89.4**	**45.7**	99.5	51.8	57.3	58.0	86.7	85.0	73.6	50.4	88.2	64.7	56.8	62.8	62.4	-
	CoPDASeg (Ours)	86.5	**45.6**	87.8	**45.1**	2.5	**54.5**	**61.6**	**58.9**	**88.9**	88.8	**76.9**	48.4	**92.2**	**83.5**	**61.8**	61.8	**65.3**	**72.5**
HRDA	HRDA (Hoyer et al., 2022b)	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	92.9	79.4	52.8	89.0	64.7	63.9	64.9	65.8	72.4
	BLV (Wang et al., 2023a)	87.6	47.9	**90.5**	50.4	6.9	57.1	64.3	**65.3**	86.9	93.4	78.9	54.9	89.1	62.9	65.2	**66.8**	66.8	73.4
	MIC (Hoyer et al., 2023)	86.6	50.5	89.3	47.9	7.8	**59.4**	66.7	63.4	87.1	94.6	81.0	58.9	90.1	61.9	67.1	64.3	67.3	74.0
	PiPa (Chen et al., 2023)	88.6	50.1	90.0	**53.8**	7.7	58.1	**67.2**	63.1	88.5	94.5	79.7	57.6	**90.8**	**70.2**	65.1	**66.9**	68.2	74.8
	CDAC (Wang et al., 2023b)	**93.1**	**68.5**	89.8	51.2	**8.9**	59.4	65.5	65.3	84.7	94.4	**81.2**	57.0	90.5	56.9	66.8	66.4	**68.7**	**75.4**
	RTea (Zhao et al., 2023)	87.8	49.0	90.3	50.3	5.5	58.6	66.0	61.4	86.8	93.1	79.5	53.1	89.5	65.1	63.7	64.6	66.5	73.0
	MICDrop (Yang et al., 2024)	82.8	42.6	**90.5**	**51.6**	**9.6**	**61.0**	65.7	65.0	**89.1**	**95.0**	**81.1**	**59.7**	90.6	68.3	**67.4**	66.5	67.9	-
	CoPDASeg (Ours)	**90.5**	**55.8**	89.9	44.5	7.4	**64.5**	**70.2**	**65.2**	**91.6**	**94.8**	80.9	**59.3**	**94.8**	**84.6**	**67.2**	65.9	**70.5**	**77.4**

Evaluation. The IoU is adopted as our evaluation metric which is broadly applied in many semantic segmentation tasks. We report the result of per-class IoU and mean IoU for all the classes. Following previous works (Jiang et al., 2022; Zhang et al., 2021), we evaluate our method on 19 common categories for GTA5 → Cityscapes and both 16 and 13 common categories for SYNTHIA → Cityscapes, respectively. The best and second-best results are highlighted in red and blue, respectively.

4.2. Comparisons with the state-of-the-arts

In order to evaluate the efficiency, we compare our CoPDASeg with two typical DASS method categories that are all based on the ResNet (He et al., 2016) backbone:

- Adversarial training methods, such as AdaptSeg (Tsai et al., 2018), ADVENT (Vu et al., 2019) and EFA (Chen et al., 2021).
- Self-training methods, such as PLCA (Kang et al., 2020), DACS (Tranheden et al., 2021), ProDA (Zhang et al., 2021) and CPSL (Li et al., 2022a).

Furthermore, we also compare our method with the Transformer-based methods, such as DAformer (Hoyer et al., 2022a), HRDA (Hoyer et al., 2022b), BLV (Wang et al., 2023a), MIC (Hoyer et al., 2023), CDAC (Wang et al., 2023b), RTea (Zhao et al., 2023) and MICDrop (Yang et al., 2024). We evaluate our proposed PatchMix method against several competitive UDA methods on the GTA5 → Cityscapes and SYNTHIA → Cityscapes benchmarks, and report results.

Results on GTA5 → Cityscpaes. We compare the proposed CoPDASeg with previous DASS methods (Chen et al., 2021; Zhang et al., 2021; Wang et al., 2023a; Hoyer et al., 2023) in Table 1. We have the following observations:

- CoPDASeg based on DeepLabV2 achieves 68.1% mIoU, outperforming the baseline model trained merely on source data by a large margin of 29.5% mIoU. Compared with the previous adversarial training method EFA (Chen et al., 2021), CoPDASeg improves the mIoU by 19.3%. CoPDASeg surpasses the second-best method MIC (Hoyer et al., 2023) by 3.9% mIoU.
- CoPDASeg obtains significant improvements in many hard classes. For example, CoPDASeg achieves 44.1% mIoU in the "train" class. Many previous methods (Zhang et al., 2021; Wang et al.,

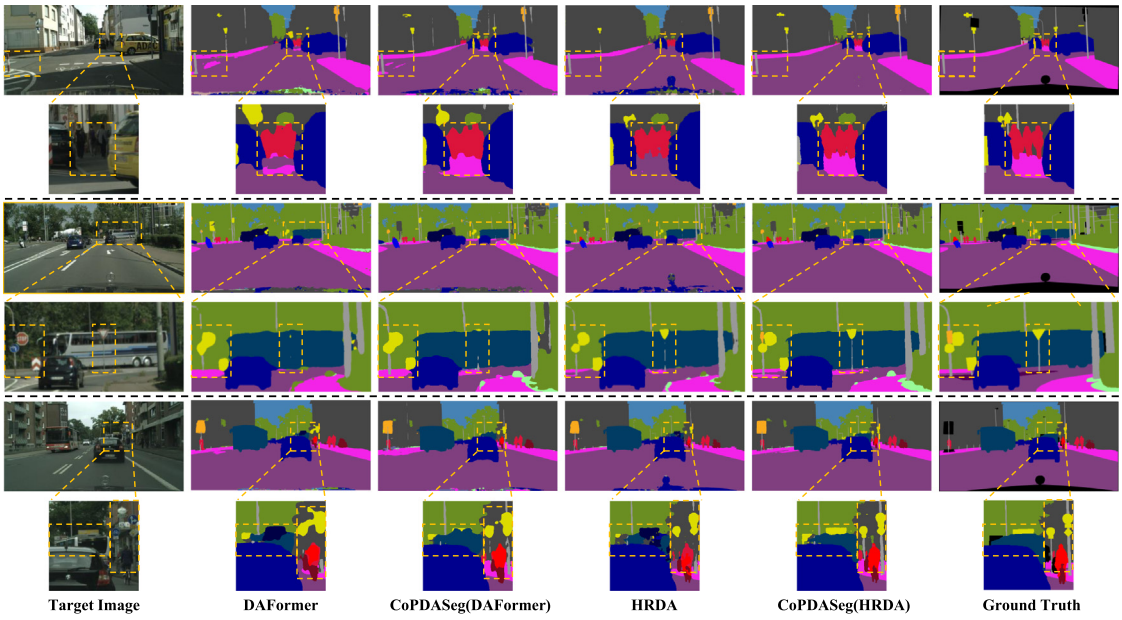


Fig. 4. Qualitative segmentation results for GTA5 (Richter et al., 2016) → Cityscapes (Cordts et al., 2016). From left to right: target image, the segmentation results predicted by DAFormer (Hoyer et al., 2022a), CoPDASeg (DAFormer), HRDA (Hoyer et al., 2022b), CoPDASeg (HRDA), and Ground Truth. We deploy the gold dash boxes to highlight different prediction parts.

2023a; Hoyer et al., 2023) fail to predict the "train" class well because this class is rarely presented in an image and has a significantly different appearance across domains. It demonstrates that CoPDASeg indeed helps the model to produce more robust features against varying environments and effectively deals with the domain gap, especially the hard classes, such as "train", "motorcycle" and "bike".

We further apply CoPDASeg to the Transformer-based architectures (DAFormer Hoyer et al., 2022a and HRDA Hoyer et al., 2022b) and compare with the other Transformer-based DASS methods, such as BLV (Wang et al., 2023a), MIC (Hoyer et al., 2023) and CDAC (Wang et al., 2023b) to demonstrate the effectiveness of CoPDASeg. Table 1 gives the comparison results on the task of GTA5 → Cityscpaes. We observe:

- Based on DAFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b), CoPDASeg surpasses MICDrop. Specifically, the mIoU score is increased by 2.3% and 0.5% respectively for HARD and MICHARD with the baseline of MICDrop. Similar experimental results can be observed from the comparison between CoPDASeg and MICDrop based on the DAFormer.
- Among all the 19 categories, CoPDASeg achieves the best or second-best performance in most of them and performs especially well in the hardest categories, such as "pole", "train", "bike" and so on.

It indicates that CoPDASeg is still competitive on the new Transformer-based architecture.

Results on SYNTHIA → Cityscpaes. Similar to previous methods (Tsai et al., 2018), we also report the DASS performance of the 16 and 13 common categories on the task of SYNTHIA → Cityscpaes in Table 2, with comparisons to the state-of-the-art DASS approaches (Chen et al., 2021; Hoyer et al., 2023). It can be seen that:

- Among all the 16 categories, CoPDASeg achieves the best scores in 6 categories, most of which are hard classes, *e.g.*, "light" and "motorcycle".
- CoPDASeg achieves the mIoU score by 64.4% and 72.6% over the 16 and 13 categories respectively, which outperforms the EFA (Chen et al., 2021) by 19.4% mIoU and 20.8% mIoU* over the 16 and 13 categories, respectively. CoPDASeg achieves a 70.5% mIoU score for the HRAD-based method, exceeding those of MICDrop. As for the DAFormer-based method, the mIoU score is increased by 2.9% compared to MICDrop.
- CoPDASeg obtains improvement over the second best method MIC (Hoyer et al., 2023) by 1.6% mIoU and 1.9% mIoU*.

Similarly, the comparison results of SYNTHIA → Cityscapes with the Transformer-based architectures (DAFormer Hoyer et al., 2022a, HRDA Hoyer et al., 2022b) are shown in Table 2. We have the following observations:

- CoPDASeg based on DAFormer (Hoyer et al., 2022a) achieves 65.3% and 72.5% mIoU scores over the 16 and 13 categories respectively and outperforms the second-best method RTea (Zhao et al., 2023) by a large margin of 2.3% and 3.0%.
- Based on HRDA (Hoyer et al., 2022b), CoPDASeg achieve mIoU of 70.5% and 77.4% for the two evaluation metrics, respectively. It surpasses the second-best method CDAC (Wang et al., 2023b) by 1.8% mIoU and 2.0% mIoU* over the 16 and 13 categories.
- Among all the 16 categories, CoPDASeg shows superior results in many important classes (*e.g.,* "traffic light", "bus", "motorcycle", *etc*).

These results reveal that CoPDASeg remains competitive on the task of SYNTHIA → Cityscapes.

Qualitative Results. In Fig. 4, we visualize the segmentation results of CoPDASeg based on DAFormer (Hoyer et al., 2022a) and HRDA (Hoyer et al., 2022b) on GTA5 → Cityscapes. It indicates that the results of CoPDASeg are smoother and more accurate than the baseline model, especially for the hard classes. For example, from the third row, it can be seen that CoPDASeg predicts smoother edges of both traffic signs and poles. In addition, it successfully segmented the bus behind the car. These results state the superiority of the proposed CoPDASeg over the baseline model. We think it is because CoPDASeg explicitly encourages pixel-level consistency against different contexts, which effectively enhances the robustness of edge predictions.

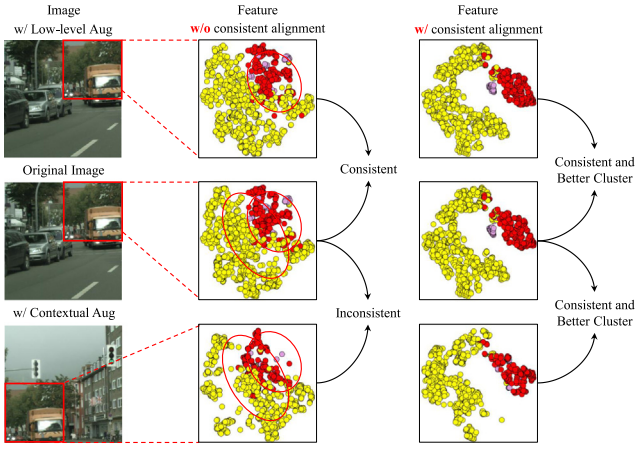


Fig. 5. Visual comparison results between contextual augmentation and low-level augmentation using t-SNE visualization for features of the overlapping region (shown in the red box). **Left**: input crops from the same image, where the first row and the third row apply the low-level and contextual augmentation respectively. **Middle**: t-SNE results of the model trained with labeled data only. Note that the three visualizations are in the same t-SNE space, and the dots with the same color represent the features of the same class. **Right**: t-SNE results of our method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

t-SNE Visualization. To better demonstrate our intuition, we draw the t-SNE visualizations of learned representations for low-level augmentation and our proposed contextual augmentation in Fig. 5. As shown in the second column of Fig. 5, it can be observed that the embedding distribution changes much more significantly under contextual augmentation (*i.e.*, second and third row of the second column) than low-level augmentations (*i.e.*, first and second row of the second column). It proves that our contextual augmentation provides stronger and more effective perturbations to promote the robustness of the model. Besides, after our proposed contrastive learning, the results (*i.e.*, second and third row of the third column) show that our method successfully aligns the features for the overlapping region and separates them for different categories. That is, it minimizes intra-class variations and maximizes inter-class variations, regardless of domains.

4.3. Ablation studies

We conduct ablation experiments to analyze the effectiveness of our proposed method. All experiments are conducted on GTA5 → Cityscapes.

Effectiveness of Each Component. In this section, we validate the contribution of each component in CoPDASeg using GTA5 → Cityscapes. For the convenience of expression, we abbreviate "pixel-to-pixel contrastive loss", "prototype-to-prototype contrastive loss" and "Non-linear Projection Head" with "$\mathcal{L}_{pixel}$", "$\mathcal{L}_{proto}$" and "$Proj$".

Table 3 shows the corresponding results by switching on each component over the baseline. CoPDASeg achieves 66.3% and 66.6% mIoU scores respectively after only using the proposed $\mathcal{L}_{pixel}$ or $\mathcal{L}_{proto}$ to achieve consistent alignment. It can be seen that either using $\mathcal{L}_{pixel}$ or $\mathcal{L}_{proto}$ to implement CoPDASeg can obtain extra gains of 2.1% mIoU and 2.4% mIoU, respectively. This demonstrates that both $\mathcal{L}_{pixel}$ and $\mathcal{L}_{proto}$ play key roles in improving the segmentation performance by accomplishing CoPDASeg at the pixel level and class level. As shown in the last two rows of Table 3, the mIoU of the adapted model decreases moderately without the non-linear projection head. It supports the importance of the projection head. By combining the proposed two contrastive loss functions with the projection head, we further obtain an improvement of 3.9% mIoU score over the baseline. The results show that each of the proposed components plays an important role in consistent alignment and effectively mitigates the domain gap.

Table 3
Ablation studies of each loss function for GTA5 → Cityscapes.

Method	$\mathcal{L}_{pixel}$	$\mathcal{L}_{proto}$	$Proj$	mIoU
CoPDASeg				64.2
	✓		✓	66.3
		✓	✓	66.6
	✓	✓		67.2
	✓	✓	✓	68.1

Table 4
Ablation studies on the key components of our proposed contextual augmentation.

Method	Source	Target	Class Mix	Low-level Aug	mIoU
Contextual Augmentation	✓			✓	65.7
		✓		✓	61.6
			✓		67.6
			✓	✓	68.1

Table 5
Ablation studies of calculating prototype on different regions.

Calculate prototype	Whole patch	Overlap region
mIoU	67.4	68.1

Ablation Studies in Contextual Augmentation Module. Table 4 lists the results of using each individual component in the contextual augmentation module. We use the source domain images or target domain images to replace the mixed images in the contextual augmentation. It leads to an obvious mIoU drop of 2.4% and 6.5% respectively. It demonstrates that achieving contextual augmentation on one domain influences subsequent cross-domain alignment processes. It can be seen that the latter brings a more serious decline in performance. Due to the lack of the target domain annotation, a large amount of noise on the pseudo label leads to insufficient learning during consistent alignment. In addition, image augmentation is used to further strengthen the difference of features between two patches and brings a performance gain of more than 0.5% mIoU. It supports our claim that a stronger perturbation plays a key role in consistency regularization.

Semantic Alignment with Different Strategies. In this section, we compare two prototype computation strategies in prototype-to-prototype contrastive learning: computing prototype from (1) the whole patch or (2) the overlapping region, respectively. The comparison results are shown in Table 5. We can observe that the first strategy is 0.7% lower than the second one, verifying the importance of location information for semantic consistent alignment. This observation indicates that the first strategy introduces more noise from the non-overlapping region of pseudo labels during consistency regularization, which harms the performance of our method.

5. Conclusion

In this paper, we aim to design stronger and more effective perturbations for cross-domain consistency learning. Thus, we propose a new augmentation method, *i.e.*, contextual augmentation, and combine it with contrastive learning from both pixel and class levels. Specifically, contextual augmentation first mixes two domain information by class mix and then randomly crops two patches with an overlapping region from the mixed image. Then, pixel-to-pixel and prototype-to-prototype contrastive learning are introduced to realize the consistency alignment at both the pixel and class levels. The effectiveness of our method is demonstrated on GTA5 → Cityscapes and SYNTHIA → Cityscapes

benchmarks. It can be observed that our method brings large improvement for hard classes, *e.g.,* "traffic light", "bus", "bike", *etc*. Besides, the architecture of our model is neat as it only adopts one single-stage pipeline without any extra special training techniques.

CRediT authorship contribution statement

Meiqin Liu: Project administration, Supervision, Formal analysis, Writing – review & editing, Investigation. **Zilin Wang:** Visualization, Formal analysis, Writing – original draft, Software, Validation. **Chao Yao:** Writing – review & editing. **Yao Zhao:** Conceptualization. **Wei Wang:** Resources, Validation. **Yunchao Wei:** Formal analysis, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62120106009, 62372036, 62332017, U22A2022, and U24B20179).

Data availability

Data will be made available on request.

References

Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C., 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8219–8228.

Araslanov, N., Roth, S., 2021. Self-supervised augmentation consistency for adapting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15384–15394.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 213–229.

Chen, Y., Hu, H., 2021. Y-Net: Dual-branch joint network for semantic segmentation. ACM Trans. Multimed. Comput. Commun. Appl. 17 (4), 0–22.

Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B., 2019a. CrDoCo: Pixel-level domain transfer with cross-domain consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1791–1800.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, T., Wang, S.H., Wang, Q., Zhang, Z., Xie, G.S., Tang, Z., 2021. Enhanced feature alignment for unsupervised domain adaptation of semantic segmentation. IEEE Trans. Multimed. 24, 1042–1054.

Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., Huang, J., 2019b. Progressive feature alignment for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 627–636.

Chen, M., Zheng, Z., Yang, Y., Chua, T.S., 2023. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1905–1914.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1290–1299.

Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 17864–17875.

Corbiere, C., Thome, N., Saporta, A., Vu, T.H., Cord, M., Perez, P., 2021. Confidence estimation via auxiliary models. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6043–6055.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. pp. 1–22.

Dundar, A., Liu, M.-Y., Yu, Z., Wang, T.-C., Zedlewski, J., Kautz, J., 2020. Domain Stylization: A fast covariance matching framework towards domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell. 43 (7), 2360–2372.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. In: Proceedings of the International Conference on Learning Representations. pp. 1–16.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In: Proceedings of the International Conference on Machine Learning. pp. 1989–1998.

Hoyer, L., Dai, D., Van Gool, L., 2022a. 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. pp. 9924–9935.

Hoyer, L., Dai, D., Van Gool, L., 2022b. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 372–391.

Hoyer, L., Dai, D., Wang, H., Van Gool, L., 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11721–11732.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. CCNet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612.

Jiang, Z., Li, Y., Yang, C., Gao, P., Wang, Y., Tai, Y., Wang, C., 2022. Prototypical contrast adaptation for domain adaptive semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 36–54.

Kang, G., Wei, Y., Yang, Y., Zhuang, Y., Hauptmann, A., 2020. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 3569–3580.

Kim, M., Byun, H., 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12975–12984.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90.

Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., Jia, J., 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1205–1214.

Li, R., Li, S., He, C., Zhang, Y., Jia, X., Zhang, L., 2022a. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11593–11603.

Li, Z., Togo, R., Ogawa, T., Haseyama, M., 2022b. Learning intra-domain style-invariant representation for unsupervised domain adaptation of semantic segmentation. Pattern Recognit. 132, 108911.

Lian, Q., Lv, F., Duan, L., Gong, B., 2019. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6758–6767.

Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei Fei, L., 2019. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 82–92.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations. pp. 1–8.

Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y., 2019. 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. pp. 2507–2516.

Mei, K., Zhu, C., Zou, J., Zhang, S., 2020. Instance adaptive self-training for unsupervised domain adaptation. In: Proceedings of the European Conference on Computer Vision. pp. 415–430.

Melas-Kyriazi, L., Manrai, A.K., 2021. PixMatch: Unsupervised domain adaptation via pixelwise consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12435–12445.

Olsson, V., Tranheden, W., Pinto, J., 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. pp. 1369–1378.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S., 2020. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3764–3773.

Richter, S.R., Vineet, V., Roth, S., Koltun, V., 2016. Playing for data: Ground truth from computer games. In: Proceedings of the European Conference on Computer Vision. pp. 102–118.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M., 2016. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3234–3243.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.-L., 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 596–608.

Song, L., Li, Y., Li, Z., Yu, G., Sun, H., Sun, J., Zheng, N., 2019. Learnable tree filter for structure-preserving feature transform. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1711–1721.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1195–1204.

Tranheden, W., Olsson, V., Pinto, J., Svensson, L., 2021. DACS: Domain adaptation via cross-domain mixed sampling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1379–1389.

Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7472–7481.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 5998–6008.

Vu, T.H., Jain, H., Bucher, M., Cord, M., 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. pp. 2517–2526.

Wang, Q., Dai, D., Hoyer, L., Van Gool, L., Fink, O., 2021a. Domain adaptive semantic segmentation with self-supervised depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8515–8525.

Wang, Y., Fei, J., Wang, H., Li, W., Bao, T., Wu, L., Zhao, R., Shen, Y., 2023a. Balancing logit variation for long-tailed semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19561–19573.

Wang, K., Kim, D., Feris, R., Betke, M., 2023b. CDAC: Cross-domain attention consistency in transformer for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11519–11529.

Wang, Y., Peng, J., Zhang, Z., 2021b. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9092–9101.

Wang, Q., Yuan, C., Liu, Y., 2019. Learning deep conditional neural network for image segmentation. IEEE Trans. Multimed. 21 (7), 1839–1852.

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L., 2021c. Exploring cross-image pixel contrast for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7303–7313.

Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q., 2020. Unsupervised data augmentation for consistency training. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 6256–6268.

Xu, S., Sun, K., Liu, D., Xiong, Z., Zha, Z.J., 2023. Synergy between semantic segmentation and image denoising via alternate boosting. ACM Trans. Multimed. Comput. Commun. Appl. 19 (2), 0–23.

Yang, L., Hoyer, L., Weber, M., Fischer, T., Dai, D., Leal-Taixé, L., Pollefeys, M., Cremers, D., Van Gool, L., 2024. Micdrop: Masking image and depth features via complementary dropout for domain-adaptive semantic segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 329–346.

Yang, Y., Soatto, S., 2020. FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4085–4095.

Yuan, Y., Fang, J., Lu, X., Feng, Y., 2019. Spatial structure preserving feature pyramid network for semantic image segmentation. ACM Trans. Multimed. Comput. Commun. Appl. 15 (3), 0–19.

Zhang, Y., David, P., Gong, B., 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2020–2030.

Zhang, L., Qi, G.J., Wang, L., Luo, J., 2019. AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2547–2555.

Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., 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. pp. 12414–12424.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.

Zhao, D., Wang, S., Zang, Q., Quan, D., Ye, X., Yang, R., Jiao, L., 2023. Learning pseudo-relations for cross-domain semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19191–19203.

Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L., 2022. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. Comput. Vis. Image Underst. 221, 103448–103448.

Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J., 2019. Confidence regularized self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5982–5991.