Improving Test-Time Efficiency in Source-Free Semantic Segmentation via Multi-Stage Self-Training

YIFANG YIN, Institute for Infocomm Research (IR), A*STAR, Singapore, Singapore
JINMING CAO, National University of Singapore, Singapore, Singapore
ZHENGUANG LIU, Zhejiang University, Hangzhou, China
GUANFENG WANG, Grabtaxi Holdings Limited, Singapore, Singapore
SHILI XIANG, Institute for Infocomm Research (IR), A*STAR, Singapore, Singapore
ROGER ZIMMERMANN, National University of Singapore, Singapore, Singapore

Source-free domain adaptive semantic segmentation aims at adapting a model trained on the source domain to the target domain without requiring access to the source data. Self-training has emerged as a leading approach to address this challenging problem. However, without robust denoising mechanisms to reduce the noise in pseudo labels, it still easily fall into biased estimates. Most existing methods address this issue by introducing novel architectures, but often at the cost of increased model complexity or reliance on additional input modalities. Different from previous studies, this article introduces *UniSFDA*, a unified multi-stage self-training framework that integrates cross-model transfer learning, uncertainty-aware pseudo label fusion, and intra-domain style augmentation, thereby enhancing both segmentation accuracy and test-time efficiency. Our proposed framework offers exceptional flexibility, with each component being independent and ready to be integrated into any existing self-training framework. Additionally, we investigate the performance of various representative segmentation models, including DeepLabv2, SegFormer, DFormer, and ViT-Adapter, within our framework. It is worth noting that *UniSFDA* is model-agnostic, allowing both source and target networks to be instantiated with arbitrary segmentation architectures, and thus readily benefiting from future advances in segmentation models. Experiments on the GTA5 → Cityscapes and SYNTHIA → Cityscapes benchmarks demonstrate the effectiveness of our framework. With DeepLabv2 (SegFormer) as the source model, *UniSFDA* establishes new state-of-the-art performance, achieving mIoU scores of 61.8% (65.4%) and 57.9% (59.6%) on the two benchmarks, respectively.

CCS Concepts: • **Computing methodologies** → **Video segmentation**; **Scene understanding**;

Additional Key Words and Phrases: Source-free domain adaptation, semantic segmentation

This research is supported by the RIE2025 Career Development Fund (Award C233312009), administered by Agency for Science, Technology and Research (A*STAR).

Authors' Contact Information: Yifang Yin (corresponding author), Institute for Infocomm Research (I^{2}R), A*STAR, Singapore, Singapore; e-mail: yin_yifang@i2r.a-star.edu.sg; Jinming Cao, National University of Singapore, Singapore, Singapore; e-mail: Jinming.ccao@gmail.com; Zhenguang Liu, Zhejiang University, Hangzhou, China; e-mail: liuzhenguang2008@gmail.com; Guanfeng Wang, Grabtaxi Holdings Limited, Singapore; e-mail: guanfeng@outlook.com; Shili Xiang, Institute for Infocomm Research (I^{2}R), A*STAR, Singapore, Singapore; e-mail: sxiang@i2r.a-star.edu.sg; Roger Zimmermann, National University of Singapore, Singapore, Singapore; e-mail: rogerz@comp.nus.edu.sg.

ACM 1551-6865/2026/3-ART102
https://doi.org/10.1145/3795886

ACM Reference format:
Yifang Yin, Jinming Cao, Zhenguang Liu, Guanfeng Wang, Shili Xiang, and Roger Zimmermann. 2026. Improving Test-Time Efficiency in Source-Free Semantic Segmentation via Multi-Stage Self-Training. *ACM Trans. Multimedia Comput. Commun. Appl.* 22, 4, Article 102 (March 2026), 20 pages.
https://doi.org/10.1145/3795886

1 Introduction

Urban scene understanding is a critical yet challenging problem in autonomous driving, as accurate perception of driving scenes is essential for vehicle navigation. Deep learning-based approaches have demonstrated strong potential, with extensive research addressing challenges such as object segmentation under both daytime [10, 36] and nighttime [37] conditions, domain-adaptive semantic segmentation [2, 45], and real-time scene parsing [61]. In this article, we focus on **Source-Free Domain Adaptation (SFDA)** for urban scenes by examining two practical limitations that frequently arise in real-world scenarios. First, the success of deep learning models usually relies on training with large-scale labeled datasets. However, in real-world applications, labeled data may be restricted from sharing due to proprietary, privacy, or profit-related concerns. This limitation compromises the feasibility of existing supervised deep learning models, as in such cases users may only have access to a pretrained source model. Second, domain shifts frequently occur in targeted application scenarios, arising from factors such as street scenes captured in cross-city [11, 19] or cross-weather [54] environment. Therefore, a pretrained model requires fine-tuning with target domain data, which are often unlabeled due to the substantial effort and cost associated with data annotation, particularly for dense prediction tasks.

Existing SFDA methods can be roughly divided into source domain generalization and target domain adaptation [32]. A key approach to addressing the challenges in source domain generalization is to construct a diverse training dataset through various data and style augmentation techniques [28, 31, 33, 62, 69, 82, 83]. Recent research has also shown that Transformer-based segmentation models typically exhibit stronger generalizability than traditional **Convolutional Neural Network (CNN)**-based segmentation models [73], particularly when trained with carefully designed training strategies [23, 25]. For target domain adaptation, previous studies largely follow a self-training paradigm, leveraging **Pseudo Labels (PLs)** generated by the source model to iteratively optimize the target model [1, 48, 80]. One critical research focus in this area is PL denoising, which aims to rectify inaccurate PLs using techniques such as uncertainty estimation [7, 40], prototypical denoising [78], stable neighbor denoising [79], and consistency preservation across perturbed inputs or modalities [45, 73]. Despite significant efforts that have been made in SFDA, most existing studies approach this problem through carefully designed architectures. Performance improvements often come at the cost of increased model complexity or reliance on additional modalities, such as joint visual–depth analysis [73]. An efficient and model-agnostic self-training framework that is designed for practical deployment and integrates all key aspects is still lacking in literature.

To bridge the gap, we present a unified two-stage pseudo-labeling and self-training framework termed *UniSFDA* that progressively trains a robust target segmentation model in a source-free setting. As shown in Figure 1, *UniSFDA* consists of a typical self-training stage and an additional intra-domain self-training stage. In the self-training stage, *UniSFDA* focuses on improving performance through transfer learning. We introduce two auxiliary networks, namely DFormer, an RGB-D segmentation model, and **Vision Transformer (ViT)**-Adapter, an adapter network built upon the Vision Transformer, to leverage the strengths of their transferable representations learned from large-scale pretraining and additional modalities such as depth information. In the intra-domain

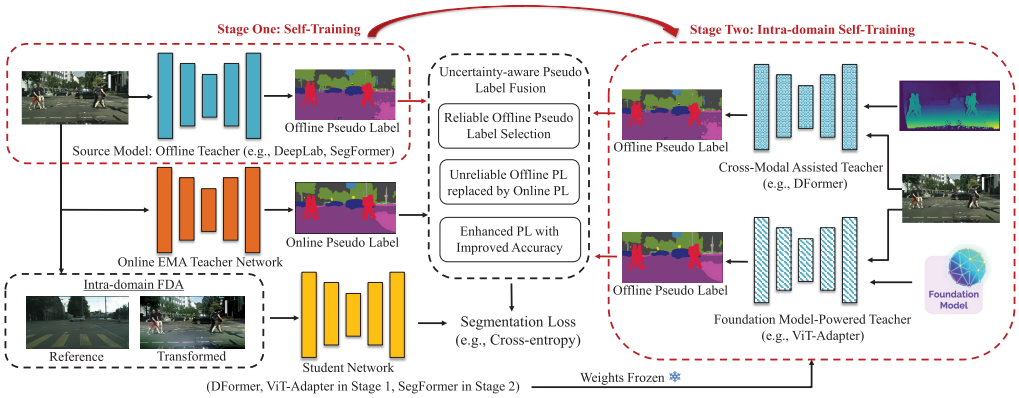

Fig. 1. System overview of our proposed unified framework for robust source-free domain adaptive semantic segmentation. Student networks trained during Stage One Self-Training (DFormer and ViT-Adapter) serve as the offline teacher models for Stage Two Intra-domain Self-Training. The student network trained during Stage Two (SegFormer) is used as the final inference model.

self-training stage, *UniSFDA* focuses on improving inference efficiency and feasibility by condensing knowledge from the large and multi-modal auxiliary networks to a single image-based target model. To enhance the quality of PLs in both stages, we propose an effective uncertainty-aware online and offline PL fusion approach. The core idea is to identify unreliable offline PLs using uncertainty measures, such as MC-Dropout [17], and replace them with more reliable online PLs generated by a mean teacher network. Additionally, we introduce and evaluate different PL aggregation strategies in the intra-domain self-training stage in conjunction with the uncertainty-aware PL fusion approach, with the objective of leveraging the complementary strengths of multiple auxiliary networks. Despite its simplicity, this fusion strategy is proven to be effective and model-agnostic, and it can work with both single and multiple teacher models, making it easily integrable into any self-training framework to achieve performance gains. Finally, we investigate the impact of intra-domain augmentations and introduce a new intra-domain **Fourier Domain Adaptation (FDA)** [70] strategy, in which reference images are randomly sampled within the same domain. Empirical studies show that this strategy is effective in both vendor-side domain generalization and client-side domain adaptation. Here, we summarize the contributions of this article as follows:

- We present a unified two-stage source-free domain adaption framework. It leverages large, cross-modal transferable auxiliary networks for enhanced segmentation while maintaining model efficiency and feasibility by condensing model size and input modalities during inference. This design achieves a balance between effectiveness and efficiency, relaxing both computational and data requirements for real-world applications.
- We propose a simple yet effective uncertainty-aware online and offline pseudo-label fusion strategy and investigate various intra-domain augmentations to enhance domain generalization and adaptation. Both components can be seamlessly integrated into existing self-training frameworks, remain agnostic to the choice of segmentation model, and thus support continuous improvement as segmentation methods advance.
- Our method effectively narrows the gap between segmentation models of different sizes and input requirements at inference. Using an image-based SegFormer model, it outperforms existing approaches by a significant margin, obtaining an **mean Intersection over Union**

(mIoU) of 61.8% (65.4%) and 57.9% (59.6%) on the Cityscapes dataset when adapting from a DeepLabv2 (SegFormer) source model trained on **Grand Theft Auto V (GTA5)** or SYNTHIA benchmarks, respectively.

The rest of this article is organized as follows. Section 2 reviews important related work on source-free domain adaptive semantic segmentation. Section 3 outlines our proposed self-training framework and details its key components. Section 4 presents the experimental setup, comparisons with state-of-the-art methods, and ablation studies. Finally, Section 5 concludes the article.

2 Related Work

2.1 Unsupervised Domain Adaptation

Unsupervised domain adaptive semantic segmentation aims to enhance model generalization in an unlabeled target domain by transferring knowledge learned from a labeled source domain [8, 9, 66]. Early approaches primarily employed adversarial training [18] to mitigate distribution discrepancies across domains [15, 42, 58]. Subsequent studies explored distribution alignment at various levels, including the image domain [22], feature representation [11], and output space [57]. More recent efforts have explored class-wise alignment strategies to achieve finer-grained feature adaptation [15, 42], or promoting the learning of intra-image pixel-level correlations and patch-wise semantic consistency across different contexts through contrastive learning [9]. Despite their effectiveness, these approaches inherently depend on the availability of source data.

2.2 Source-Free Unsupervised Domain Adaptation

Source-free unsupervised domain adaptation aims to transfer knowledge from a pretrained source model to an unlabeled target domain without access to the source data [3, 38, 50, 68, 74]. Prior research mainly addresses target domain adaptation. For instance, Teja and Fleuret [49] proposed to reduce prediction uncertainty based on feature corruption with entropy regularization. Liu et al. [38] leveraged a generator to estimate the source data distribution, based on which fake samples were synthesized for training. Among recent efforts, pseudo-labeling and self-training have emerged as predominant approaches [1, 48, 80]. Various methods have been proposed to enhance the accuracy of PLs, including uncertainty estimation with Bayesian Neural Networks [40], stable neighbor denoising [79], and consistency preservation across modalities [73]. Furthermore, Zang et al. [76] addressed a more challenging scenario and proposed a reliable knowledge propagation approach to ensure robust performance in both the source and target domains.

2.3 PL Denoising

In addition to the aforementioned methods specifically designed for SFDA [40, 73, 79], PL denoising techniques have been more broadly applied in non-SFDA and semi-supervised learning [30, 34, 44, 55]. For instance, Fang et al. [16] proposed a new data augmentation mechanism named Uncertainty-aware Patch CutMix to locate noisy PLs. Zhang et al. presented ProDA [78], which refines offline PLs based on their feature distance from the prototype of each semantic class. The mean teacher approach [56], which derives a teacher model by temporally averaging the student model's weights over training iterations, is widely adopted to produce robust online PLs. Fusion approaches have also been proposed to improve model performance, either by enforcing consistency between online and offline PLs [2, 45, 77] or by combining them to obtain enhanced training labels [47]. When access to labeled source data is restricted, adapting such PL denoising methods becomes necessary. However, their performance in the context of source-domain adaptation remains unclear.

2.4 Model Generalization

Kundu et al. partitioned SFDA into two components: (1) source domain generalization and (2) target domain adaptation [32]. Their focus has been on the former, where they developed a multi-head framework trained by augmenting the source data with diverse transformations. Hoyer et al. proposed DAFormer [23], which leverages a Transformer-based architecture as the source model to enhance generalization capability. They further extended this method to a multi-resolution framework, enabling the learning of segmentation details for small objects from high-resolution inputs [25]. While early studies primarily sought to expand the source domain through data augmentation with diverse styles [26, 28, 31, 33, 69, 82, 83], recent approaches address this challenge from a complementary perspective by harnessing the superior generalization capability of various foundation models [4, 64, 65]. In a source-free DA framework, a domain-generalized source model can produce more accurate PLs, thereby improving segmentation performance in conjunction with PL denoising approaches on the target domain.

3 Approach

We propose a unified two-stage pseudo-labeling framework for the progressive training of a robust source-free domain adaptive semantic segmentation model, as illustrated in Figure 1. In the self-training stage, *UniSFDA* focuses on improving performance by taking advantage of additional input (e.g., depth information) and foundation models. In the intra-domain self-training stage, *UniSFDA* enhances inference efficiency by condensing knowledge to a single image-based target model. In the rest of this section, we begin with a problem formulation, followed by the technical details of our proposed framework.

3.1 Problem Formulation

Source-free domain adaptive semantic segmentation aims at transferring knowledge from a source model $h^s(x)$ to a target model $h^t(x)$ using an unlabeled target dataset. Let $\mathcal{X} = \{(x_i, d_i)\}_{i=1}^n$ denote the target dataset where (x_i, d_i) represent the RGB and the optional depth modality of the ith sample, respectively. We tackle this problem by initially adapting the source model $h^s(x)$ to multiple auxiliary target models, such as $h^{a_1}(x, d)$ and $h^{a_2}(f_{FM}(x))$, by harnessing the strengths of additional modalities d or foundation models f_{FM} such as ViT. In the next stage, the auxiliary models act as teacher networks, distilling their knowledge into a single unimodal target model, such as SegFormer, thereby improving test-time efficiency while maintaining high segmentation accuracy.

3.2 Source-Free UDA via Self-Training

We begin by enhancing the robustness of pseudo-labeling and self-training through an uncertainty-aware fusion approach that integrates offline and online PLs. Next, we extend this approach in a two-stage cross-model transfer learning framework, as described in Section 3.3. In particular, multi-modal and foundation segmentation models are introduced as auxiliary teacher networks. Subsequently, PL aggregation strategies are proposed to combine their complementary strengths, enabling the uncertainty-aware fusion approach to work with multiple teacher models.

3.2.1 Offline PL Generation. In SFDA, the offline PLs are generated by feeding target images to the source model. Directly using offline PLs for training is ineffective due to the high noise introduced by the distribution mismatch between the source and target domains. A common approach is to select the most confident top percentage of predictions per class across the entire training set, forming a balanced and reliable set of PLs for self-training [32, 35, 73]. Although this approach has been shown to be effective to some extent, many of the selected predictions remain incorrect due to the poor calibration of the source model [20].

To solve this problem, we select reliable PLs based on both uncertainty and confidence [52]. The MC-Dropout [17] is adopted as an uncertainty measure to compute the variance of five stochastic forward passes with dropout enabled. Let $p^{(i,k)}$ and $u^{(i,k)}$ represent the confidence and uncertainty of pixel $x^{(i)}$ belonging to the kth class, respectively. We first derive the hard PL $y_{off}^{(i,k)}$ as:

$$y_{off}^{(i,k)} = \begin{cases} 1 & \text{if } k = \arg\max_{k'} p^{(i,k')} \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

Next, we introduce two thresholds σ_k and τ_k for each class to select hard PLs based on confidence and uncertainty, respectively. To maintain a balanced training set across classes, we set σ_k to the m_c-percentile after ranking all samples belonging to class k (i.e., $y_{off}^{(i,k)} = 1$) in descending order based on their confidence scores, and set τ_k to the m_u-percentile after ranking all samples belonging to the same class (i.e., $y_{off}^{(i,k)} = 1$) in descending order based on their uncertainty scores. Finally, we compute $g^{(i,k)}$ as:

$$g^{(i,k)} = \mathbb{1}[p^{(i,k)} \geq \sigma_k] \cdot \mathbb{1}[u^{(i,k)} \geq \tau_k], \tag{2}$$

to indicate whether the offline PL for pixel $x^{(i)}$ is selected (i.e., $g^{(i,k)} = 1$) or not (i.e., $g^{(i,k)} = 0$).

3.2.2 Online PL Generation. The online PLs are typically generated by a mean teacher network, which averages the student model's weights over multiple training steps [56]. A common line of research focuses on introducing input perturbations through data augmentations and enforcing consistency regularization between the teacher's output and the student's output [2, 45] or on rectifying the offline PLs based on online class prototypes [73, 78]. However, such methods usually introduce additional losses and computational cost. Let $y_{on}^{(i,k)}$ denote the hard PL generated by the online teacher network. We present a simple and effective strategy for fusing online and offline PLs, which can be optimized using a straightforward cross-entropy loss.

3.2.3 Uncertainty-Aware Online and Offline PL Fusion. Combining the strength of online and offline PLs has proven effective for non-source-free domain adaptive semantic segmentation when labeled source data is accessible [47, 77], but remains underexplored in the source-free setting. To bridge the gap, we propose first training the student model using only reliable offline PLs during the warm-up stage, and then replacing the masked unreliable offline PLs with their online counterparts. Let $y_{war}^{(i,k)}$ and $y_{fus}^{(i,k)}$ denote the PLs used in the warm-up stage and the subsequent stage, respectively, we have:

$$\begin{aligned} y_{war}^{(i,k)} &= g^{(i,k)} \cdot y_{off}^{(i,k)} \\ y_{fus}^{(i,k)} &= g^{(i,k)} \cdot y_{off}^{(i,k)} + (1 - g^{(i,k)}) \cdot y_{on}^{(i,k)}. \end{aligned} \tag{3}$$

By substituting Equation (2) into Equation (3), we have:

$$\begin{aligned} y_{war}^{(i,k)} &= \mathbb{1}[p^{(i,k)} \geq \sigma_k] \cdot \mathbb{1}[u^{(i,k)} \geq \tau_k] \cdot y_{off}^{(i,k)} \\ y_{fus}^{(i,k)} &= \mathbb{1}[p^{(i,k)} \geq \sigma_k] \cdot \mathbb{1}[u^{(i,k)} \geq \tau_k] \cdot y_{off}^{(i,k)} \\ &\quad + (1 - \mathbb{1}[p^{(i,k)} \geq \sigma_k] \cdot \mathbb{1}[u^{(i,k)} \geq \tau_k]) \cdot y_{on}^{(i,k)}, \end{aligned} \tag{4}$$

where $y_{war}^{(i,k)} = 0$ or $y_{fus}^{(i,k)} = 0$ indicates that the pixel is masked and will not be used in the loss calculation.

3.2.4 Objectives. Given the fused PLs, segmentation models like SegFormer [67] can be effectively optimized by using cross-entropy loss as the main objective:

$$\ell_{ce}(\hat{y}, p) = -\sum_{i=1}^{H\times W}\sum_{k=1}^{K}\hat{y}^{(i,k)}\log p^{(i,k)}. \tag{5}$$

where $\hat{y}$ denotes either y_{war} or y_{fus} depending on the stage, and p is the softmax probability output by the student model.

3.3 Two-Stage Cross-Model Transfer Learning

While our uncertainty-aware online and offline PL fusion approach generally enhances self-training performance, the potential of transfer learning from foundation models remains underexplored in source-free domain adaptive semantic segmentation. Recent studies have utilized foundation models to improve the generalizability of semantic segmentation models [4, 65]. When employed as the source model, they can generate more accurate PLs and subsequently enhance domain adaptation performance. Comparatively, in this section, we explore a complementary challenge: How to harness transfer learning to enhance domain adaptation from a predefined source model more effectively?

We propose a two-stage cross-model transfer learning framework to address this challenge. As illustrated in Figure 1, our approach consists of a self-training stage, where auxiliary models generate more accurate PLs, followed by an intra-domain self-training stage, which consolidates their knowledge into a single unimodal model for improved efficiency.

3.3.1 Auxiliary Model Selection. We select two representative auxiliary models, namely a cross-modal segmentation model and a ViT-based segmentation model, and evaluate their effectiveness within our framework.

DFormer. This model is a novel RGB-D framework pretrained to learn transferable representations for RGB-D segmentation tasks [72]. Previous studies have demonstrated that incorporating depth as an additional input can achieve substantial performance gain [59, 60, 73]. Without the need for additional annotation, depth information can be easily derived from stereo images or video sequences using self-supervised depth estimation models for stereoscopic depth [54, 60] or monocular depth [63]. Thus, we adopt DFormer to exploit its cross-modal transferable representations, aiming to enhance performance in the target domain.

ViT-Adapter. This model adapts the ViT [14] to dense prediction tasks such as semantic segmentation [12]. It employs a standard ViT as the backbone to capture robust representations from large-scale multi-modal data. When transferring to downstream tasks, a pretraining-free adapter is incorporated to inject image-related inductive biases into the model, optimizing its performance for tasks such as semantic segmentation of urban scenes in our case. The adapter has a small number of trainable parameters. However, a potential drawback is that the superior transferability offered by ViT-Adapter-L, with its overall larger size, makes it bulkier compared to typical semantic segmentation models like SegFormer [67].

3.3.2 One to Multi Self-Training. In this one-to-multi self-training stage, we follow the procedure outlined in Section 3.2, starting by generating offline PLs using the source model. Next, we adopt either DFormer or ViT-Adapter as the student model and optimize them individually using the uncertainty-aware online and offline PL fusion strategy defined in Equation (4). Equipped with the robust and transferable representations captured in their pretrained weights, these models also exhibit improved performance in the target domain. However, the feasibility of such models in practical applications is constrained by additional input requirements, such as depth, or by their

large model size. Therefore, we further introduce a multi-to-one intra-domain self-training stage to alleviate these requirements during inference.

3.3.3 Multi to One Intra-Domain Self-Training. In this multi-to-one intra-domain self-training stage, the auxiliary models trained in the previous self-training stage are subsequently used as the offline teacher to generate offline PLs. We refer to this stage as intra-domain self-training as both the teacher models and the student model are learned from unlabeled data within the same target domain. Furthermore, to better exploit the complementary strengths of multiple teacher models, we introduce three PL aggregation strategies, namely *Ensemble*, *Ensemble (Intersection)*, and *Ensemble (Union)*, the details of which are presented later.

Ensembles. Let $p_D^{(i,k)}$, $u_D^{(i,k)}$, $p_V^{(i,k)}$, and $u_V^{(i,k)}$ represent the confidence and uncertainty of pixel $x^{(i)}$ belonging to the kth class generated by DFormer and ViT-Adapter, respectively. This strategy computes the aggregated $\tilde{p}^{(i,k)}$ and $\tilde{u}^{(i,k)}$ by taking the average of their ensembles using Equation (6), and then computes the hard PLs $y^{(i,k)}$[1] in Equation (4) by setting $p^{(i,k)} = \tilde{p}^{(i,k)}$ and $u^{(i,k)} = \tilde{u}^{(i,k)}$. This approach is adopted by default in our experiments unless otherwise specified:

$$\begin{aligned} \tilde{p}^{(i,k)} &= \frac{1}{2} \cdot (p_D^{(i,k)} + p_V^{(i,k)}) \\ \tilde{u}^{(i,k)} &= \frac{1}{2} \cdot (u_D^{(i,k)} + u_V^{(i,k)}) \end{aligned}. \tag{6}$$

Ensembles (Intersection). This strategy first computes the hard PLs $y_D^{(i,k)}$ and $y_V^{(i,k)}$ using Equation (4) for DFormer and ViT-Adapter separately, based on their respective confidence $p_D^{(i,k)}$ and $p_V^{(i,k)}$ and uncertainty $u_D^{(i,k)}$ and $u_V^{(i,k)}$. The resulting hard PLs are then aggregated via their intersection according to Equation (7):

$$y_{Intersection}^{(i,k)} = \begin{cases} y_D^{(i,k)}, & \text{if } y_D^{(i,k)} = y_V^{(i,k)}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Ensembles (Union) . This strategy first computes the hard PLs $y_D^{(i,k)}$ and $y_V^{(i,k)}$ using Equation (4) for DFormer and ViT-Adapter separately, based on their respective confidence $p_D^{(i,k)}$ and $p_V^{(i,k)}$ and uncertainty $u_D^{(i,k)}$ and $u_V^{(i,k)}$. The resulting hard PLs are then aggregated via their union according to Equation (8):

$$y_{\text{Union}}^{(i,k)} = \begin{cases} y_D^{(i,k)}, & \text{if } y_D^{(i,k)} = y_V^{(i,k)} \text{or } y_V^{(i,k)} = 0, \\ y_V^{(i,k)}, & \text{else if } y_D^{(i,k)} = 0. \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

To summarize, the default strategy, *Ensembles*, aggregates the confidence and uncertainty scores prior to generating the final hard PLs, while the other two strategies first derive hard PLs from individual teacher models and then perform aggregation directly on the hard PLs. The performance of different PL aggregation strategies will be compared and discussed in the experimental section.

For the student model in this stage, we adopt SegFormer due to its excellent performance and simplicity [23, 67]. It is optimized using one of the PL aggregation approach in conjunction with uncertainty-aware PL fusion, and serves as the final segmentation model used for inference. Considering that the depth information may not always be available during test-time inference, our method improves the model's feasibility by relaxing this requirement and reducing the model size.

[1]We omit the subscript war or fus for simplicity.

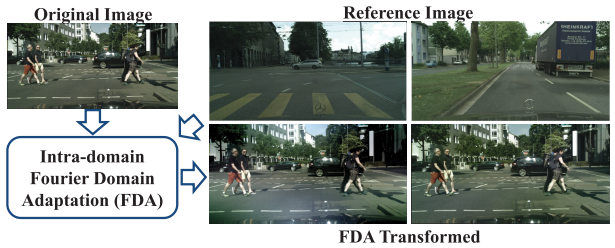

Fig. 2. Examples of images transformed using intra-domain FDA.

3.4 Intra-Domain Style Augmentation

Recent studies have demonstrated that Transformer-based segmentation models [23, 25] exhibit significantly better generalizability than traditional DeepLabv2 [32] when trained with carefully designed training strategies. Following this trend, we would like to investigate the extent to which the performance of a SegFormer can be enhanced using only simple and non-learning-based intra-domain style augmentations. In addition to the commonly used geometric distortion and photometric distortion, we also explore the following three efficient augmentation strategies.

CutMix [75] *and ClassMix* [46]. Both augmentations combine two images by cutting a region from one image and pasting it onto another. The labels of the mixed image are also interpolated based on the area of the pasted region. While the former cuts a random rectangular region, the latter selects a subset of object classes from one image and pastes them onto another image.

Intra-Domain FDA [70]. FDA was initially proposed for unsupervised domain adaptation, transforming source images based on the style of target images. As shown in Figure 2, we extend this idea by randomly sampling an image within the same domain as the reference style and demonstrate its benefits for both source domain generalization and target domain adaptation.

Weather and Cartoon Style Augmentation [29]. Following previous work [32], we apply weather and cartoon augmentations from imgaug to enhance the diversity of source images by simulating different environmental conditions or artistic styles, thereby improving the generalizability of the source model.

4 Experiments

4.1 Experimental Settings

Dataset. We evaluate our proposed method by adapting from GTA5 [51] and SYNTHIA [53] to the Cityscapes [13]. GTA5 dataset contains 24,966 annotated images with 1,914 × 1,052 resolution taken from the game GTA5, while SYNTHIA contains 9,400 annotated images with 1,280 × 760 resolution generated using a virtual city created with the Unity development platform. Both GTA5 and SYNTHIA are synthetic datasets generated through simulation, providing highly accurate pixel-level semantic annotations. They are commonly used as supplementary datasets to improve segmentation performance on real-world data. The Cityscapes dataset contains real-world street-scene images captured from vehicles at a resolution of 2,048 × 1,024, which is further divided into 2,975 training images and 500 validation images.

Evaluation Metric. We report the **Intersection over Union (IoU)** on the 19 common categories shared by GTA5 and Cityscapes and the 16 common categories shared by SYNTHIA and Cityscapes. Following previous studies, we also report the results on 13 of the 16 common categories shared by the SYNTHIA and Cityscapes datasets.

The best score for each column is highlighted. *UniSFDA* and *UniSFDA** indicate that DeepLabv2 and SegFormer are used as the source models, respectively. mIoU and mIoU* denote the averaged scores across 16 and 13 categories, respectively.

Implementation Details. We conduct experiments using two source models. To ensure a fair comparison with existing methods, we used the pretrained models on GTA5 and SYNTHIA provided by GtA [32], which are based on the DeepLabv2 [6] architecture with ResNet-101 [21] as the backbone. Additionally, we trained two Transformer-based segmentation models, i.e., SegFormer with MiT-B5 [67], on GTA5 and SYNTHIA, respectively, using the intra-domain style augmentations described in Section 3.4. For the self-training of DFormer, ViT-Adapter, and SegFormer, we adhere to their respective default training settings, set the mini-batch size to 4, and train on 4 NVIDIA L40 GPUs for up to 60 epochs. The stereo depth [54, 60] was used as the input to DFormer and the large-sized BEiT model pretrained on ImageNet-22k was adopted as the backbone in ViT-Adapter. For uncertainty-aware PL generation, we set both m_c and m_u to 66% by default when computing the per-class thresholds σ_k and τ_k in Equation (2). A sensitivity analysis on the impact of these two parameters is presented in Section 4.3.

4.2 Comparison with State-of-the-Art Methods

We compared our proposed method with the state-of-the-art source-free domain adaptive semantic segmentation models in Tables 1 and 2. $UniSFDA_S$, $UniSFDA_D$, and $UniSFDA_V$ denote the SegFormer, DFormer, and ViT-Adapter that are adopted as the student model in the first self-training stage, all trained using DeepLabv2 as the source model. $UniSFDA_{E \rightarrow S}$ represents the SegFormer student model trained in the second intra-domain self-training stage by transferring knowledge from the two auxiliary models, i.e., $UniSFDA_D$ and $UniSFDA_V$, obtained in the first stage. When our method is marked with an asterisk ($*$), e.g., $UniSFDA_V{}^*$, it indicates that the model is adapted from a SegFormer source model instead of a DeepLabv2 source model.

The experimental results show that the two auxiliary models we selected can effectively improve the adaptation performance, benefiting from the transferable representations learned from large-scale pretraining. For example, $UniSFDA_V$ ($UniSFDA_V{}^*$) outperformed $UniSFDA_S$ ($UniSFDA_S{}^*$) by a significant 2.4% (4.6%) on GTA5 $\rightarrow$ Cityscapes, which demonstrates the effectiveness of our proposed cross-model transfer learning strategy. Moreover, with the application of the second intra-domain self-training phase, the mIoU was further improved to 61.8% on GTA5 $\rightarrow$ Cityscapes and 57.9% on SYNTHIA $\rightarrow$ Cityscapes, establishing new state-of-the-art results with the DeepLabv2 source model. Lastly, by replacing the DeepLabv2 source model with our SegFormer source model trained purely based on simple and efficient intra-domain augmentations, we achieved the best mIoU of 65.4% on GTA5 $\rightarrow$ Cityscapes and 59.6% on SYNTHIA $\rightarrow$ Cityscapes, obtaining an absolute performance gain of 1.7–3.2%.

4.3 Ablation Study and Discussion

Model Justification in Self-Training Stage. We evaluate the key components of the *Stage One Self-Training* in our proposed framework and report the results in Table 3. As it shows, we start with a DeepLabv2 source model that achieves an mIoU of 44.0% on GTA5 $\rightarrow$ Cityscapes, and transfer its knowledge to SegFormer, DFormer, and ViT-Adapter, respectively, to study the impact of using foundation models. By comparing different student models, we observed that DFormer and ViT-Adapter generally outperformed SegFormer, which can be attributed to their superior transferability from large-scale pretraining. Next, we investigate the effectiveness of the proposed online and offline PL fusion approach. The results in the second, fifth, and eighth rows were obtained using models trained with only offline PLs, which resulted in limited performance. By incorporating online PLs to replace unreliable offline PLs, we observed a significant performance gain of 3.5–4.8% across

Table 1. Per-Class IoU (%) and mIoU (%) Comparison of GTA5 → Cityscapes Adaptation

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vege.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
SFDA [38]	84.2	39.2	82.7	27.5	22.1	25.9	31.1	21.9	82.4	30.5	85.3	58.7	22.1	80.0	33.1	31.5	3.6	27.8	30.6	43.2
URMA [49]	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
LD [74]	91.6	53.2	80.6	36.6	14.2	26.4	31.6	22.7	83.1	42.1	79.3	57.3	26.6	82.1	41.0	50.1	0.3	25.9	19.5	45.5
DTAC [68]	78.0	29.5	83.0	29.3	21.0	31.8	38.1	33.1	83.8	39.2	80.8	61.0	30.0	83.9	26.1	40.4	1.9	34.2	43.7	45.7
SRDA [3]	90.5	47.1	82.8	32.8	28.0	29.9	35.9	34.8	83.3	39.7	76.1	57.3	23.6	79.5	30.7	40.2	0.0	26.6	30.9	45.8
AUGCO [48]	92.6	**84.6**	86.8	84.8	**56.1**	**82.2**	45.3	**63.4**	43.8	28.7	26.8	14.0	41.1	34.9	16.7	29.6	**45.7**	5.6	12.6	47.1
SFUDA [71]	95.2	40.6	85.2	30.6	26.1	35.8	34.7	32.8	85.3	41.7	79.5	61.0	28.2	86.5	41.2	45.3	15.6	33.1	40.0	49.4
IAPC [5]	90.9	36.5	84.4	36.1	31.3	32.9	39.9	38.7	84.3	38.6	87.5	58.6	28.8	84.3	33.8	49.5	0.0	34.1	47.6	49.4
HCL [27]	92.6	54.6	82.8	33.2	26.2	39.8	38.1	31.9	84.5	38.6	85.3	61.3	30.2	85.4	33.1	41.6	14.4	27.3	44.0	49.7
CROTS [41]	92.0	52.4	85.9	37.3	35.8	34.6	42.2	38.4	86.9	45.6	**91.1**	65.1	36.1	87.3	41.6	51.1	0.0	41.4	56.2	53.7
GtA [32]	91.7	53.4	86.1	37.6	32.1	37.4	38.2	35.6	86.7	48.5	89.9	62.6	34.3	87.2	51.0	50.8	4.2	42.7	53.9	53.4
UASFDA [40]	94.4	64.5	86.6	42.3	28.4	38.0	43.8	45.3	86.7	47.0	90.0	62.6	32.2	84.9	36.1	46.0	0.0	38.1	53.8	53.7
STvM [1]	92.1	55.1	87.6	45.5	33.3	36.1	41.8	48.6	87.9	**51.1**	87.6	63.2	8.6	87.0	47.8	59.8	0.0	50.1	60.4	54.9
DT-ST [80]	93.5	57.6	84.7	36.5	25.2	33.4	44.7	36.7	86.8	42.8	81.3	62.3	37.2	88.1	48.7	50.6	35.5	48.3	59.1	55.4
SDA [79]	93.0	54.0	84.6	35.6	30.3	31.0	41.9	41.6	87.6	44.6	86.4	62.6	38.5	87.5	48.7	42.9	36.6	49.5	58.7	55.6
Uni-UVPT [43]	93.4	53.9	87.2	43.5	35.5	39.2	43.9	34.6	87.8	50.3	85.8	65.4	33.4	88.5	53.7	61.9	32.3	39.0	51.3	56.9
CrossMatch [73]	94.5	65.5	87.4	45.7	42.6	42.3	46.7	54.5	88.3	48.0	84.7	66.0	33.4	89.9	53.5	56.8	0.0	46.9	49.4	57.7
RKP [76]	**95.9**	62.2	86.9	39.6	36.9	38.7	44.4	49.0	**89.6**	46.7	90.8	66.0	**41.4**	90.3	56.0	45.3	34.5	50.5	62.1	59.3
UniSFDA$_S$	91.7	54.1	86.4	51.3	12.6	37.8	49.1	44.7	85.0	45.7	90.3	66.4	34.2	87.4	59.8	52.6	0.0	45.7	56.7	55.4
UniSFDA$_D$	93.0	58.7	87.2	46.3	37.5	41.0	47.5	38.5	87.5	46.7	85.0	62.8	34.1	90.1	54.5	59.3	0.0	46.8	59.9	56.6
UniSFDA$_V$	93.5	61.3	**89.4**	50.6	36.6	41.6	53.8	48.0	88.3	49.3	90.9	**67.8**	35.8	88.2	39.3	57.6	0.0	45.5	59.8	57.8
UniSFDA$_{E\to S}$	94.3	64.5	89.2	**54.9**	45.1	41.3	**56.0**	53.6	88.8	50.2	90.4	66.8	38.2	**90.9**	**67.0**	**67.6**	0.0	**50.6**	**65.3**	**61.8**
*UniSFDA$_S$**	90.6	51.8	88.1	50.6	44.9	49.3	56.1	46.5	89.3	48.6	87.5	72.2	38.5	90.7	62.6	24.4	11.8	55.1	61.5	59.0
*UniSFDA$_D$**	90.4	52.0	89.1	46.7	44.1	51.3	60.1	44.4	89.9	46.2	91.4	72.1	40.6	91.4	58.0	61.9	40.3	42.4	64.5	61.9
*UniSFDA$_V$**	88.8	50.9	89.7	58.4	34.0	53.5	63.0	46.7	90.0	52.1	92.9	72.7	40.0	91.2	56.2	72.2	47.9	49.7	58.4	63.6
*UniSFDA$_{E\to S}$**	89.8	53.5	88.8	50.9	49.0	48.7	57.5	44.4	89.6	51.7	89.2	71.0	39.2	92.5	71.3	76.8	55.3	55.6	67.7	65.4

UniSFDA and UniSFDA* indicate that DeepLabv2 and SegFormer are used as the source models, respectively. The best score in each column is highlighted: bold values correspond to results using the DeepLabv2 source model, while underlined values correspond to results using the SegFormer source model.

models. Finally, we show that self-training can also benefit from the intra-domain FDA, leading to an absolute mIoU improvement of 0.5–1.1% across models. In summary, all three aspects, i.e., the use of foundation models, uncertainty-aware PL fusion, and intra-domain FDA augmentation, contribute to the overall performance gains, with Table 3 quantifying the improvements brought by each component.

Model Justification in Intra-Domain Self-Training Stage. In the intra-domain self-training stage, we transfer knowledge from stage-1 models or their ensembles to a single unimodal SegFormer, aiming to improve model efficiency and feasibility during inference. Table 4 illustrates the importance of selecting reliable PLs based on both confidence and uncertainty scores (SegFormer ⟶ SegFormer) and the effectiveness of our proposed cross-model transfer learning from auxiliary models (DFormer ⟶ SegFormer and ViT-Adapter ⟶ SegFormer). Specifically, incorporating uncertainty into PL selection improves the mIoU from 56.6 to 58.2, demonstrating the effectiveness of our approach. Furthermore, we evaluate the performance of different ensemble approaches that combine the strengths of multiple teacher models. In addition to the proposed Ensembles, Ensembles (Intersection), and Ensembles (Union), we also report the results achieved by Late Fusion, where the outputs of multiple teacher models are fused at the prediction level without performing a second round of self-training. We observe that self-training based on the ensemble of DFormer and ViT-Adapter using Equation (6) (Ensembles ⟶ SegFormer) achieved the best mIoU of 61.8%, outperforming its competitors by 1.5–3.9%.

Impact of Intra-Domain Augmentations on Model Generalization. As mentioned above, in addition to the DeepLabv2 source model, we trained a SegFormer source model to harness the generalization capability of Transformers over CNN models. According to Kundu et al. [32], efforts on source-free domain adaptive semantic segmentation can be divided into (1) vendor-side domain generalization, and (2) client-side domain adaptation. Here, we focus on vendor-side evaluation, whose objective

Table 2. Per-Class IoU (%) and mIoU (%) Comparison of SYNTHIA → Cityscapes Adaptation

Method	Road	Sidewalk	Building	Wall*	Fence*	Pole*	Light	Sign	Vege.	Sky	Person	Rider	Car	Bus	Motor	Bike	mIoU	mIoU*
SFDA [38]	81.9	44.9	81.7	4.0	0.5	26.2	3.3	10.7	86.3	89.4	37.9	13.4	80.6	25.6	9.6	31.3	39.2	45.9
AUGCO [48]	87.2	**80.1**	80.8	**77.9**	**44.5**	**82.4**	26.5	**54.8**	0.3	8.2	34.6	12.6	8.7	18.8	8.2	6.0	39.5	45.9
URMA [49]	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
DTAC [68]	77.5	37.4	80.5	13.5	1.7	30.5	24.8	19.7	79.1	83.0	49.1	20.8	76.2	12.1	16.5	46.1	41.8	47.9
LD [74]	77.1	33.4	79.4	5.8	0.5	23.7	5.2	13.0	81.8	78.3	56.1	21.6	80.3	49.6	28.0	48.1	42.6	50.1
SFUDA [71]	90.9	45.5	80.8	3.6	0.5	28.6	8.5	26.1	83.4	83.6	55.2	25.0	79.5	32.8	20.2	43.9	44.2	51.9
IAPC [5]	68.5	29.2	82.0	10.9	1.2	28.7	22.3	29.1	82.8	85.3	60.5	19.3	83.1	42.5	32.2	47.7	45.3	52.7
HCL [27]	86.7	38.1	82.7	10.0	0.6	30.3	25.4	29.7	82.8	85.9	61.9	24.8	84.5	38.9	22.6	37.9	46.4	54.0
STvM [1]	71.6	31.1	84.2	0.0	0.0	36.7	36.0	38.4	85.4	87.8	55.4	23.5	85.9	47.8	29.1	53.6	47.9	56.1
DT-ST [80]	88.9	45.8	83.3	13.7	0.8	32.7	31.6	20.8	85.7	82.5	64.4	27.8	88.1	50.9	37.6	57.3	50.7	58.8
CROTS [41]	89.4	41.6	82.7	15.1	1.2	34.7	33.7	25.7	83.7	87.9	66.6	34.6	85.4	45.9	43.5	49.6	51.3	59.3
GtA [32]	90.5	50.0	81.6	13.3	2.8	34.7	25.7	33.1	83.8	89.2	66.0	**34.9**	85.3	53.4	46.1	46.6	52.0	60.1
UASFDA [40]	93.0	59.0	83.8	16.7	1.2	35.1	36.9	35.2	84.2	89.1	64.9	34.3	82.3	38.8	41.4	52.9	53.0	61.2
Uni-UVPT [43]	88.3	43.6	84.4	32.6	2.8	40.3	43.4	29.6	83.5	87.5	**67.5**	**34.9**	85.6	49.8	34.0	53.8	53.8	60.4
SDA [79]	88.1	47.4	80.1	28.1	32.2	34.9	33.6	41.3	83.3	86.7	59.9	27.2	86.7	48.1	36.2	52.5	54.1	59.3
CrossMatch [73]	91.5	56.3	85.9	37.9	9.2	42.1	42.6	47.6	87.2	86.1	64.5	23.3	89.3	**64.5**	45.0	47.7	57.5	64.0
RKP [76]	91.6	51.0	85.2	39.0	25.2	31.6	41.0	45.4	**89.0**	88.9	63.6	33.5	**88.9**	37.9	49.1	**60.9**	57.8	63.8
UniSFDA$_S$	92.4	55.8	84.3	13.3	1.5	35.8	40.8	32.0	83.1	87.4	64.1	30.2	87.4	46.1	47.8	50.2	53.3	61.6
UniSFDA$_D$	92.3	54.0	84.7	14.0	0.8	37.8	38.4	32.9	85.7	89.6	59.2	29.7	86.5	51.0	46.8	58.7	53.9	62.3
UniSFDA$_V$	92.8	56.0	**86.2**	28.5	3.9	37.3	48.1	40.3	85.8	**89.7**	61.3	20.4	87.8	37.5	40.5	55.8	54.5	61.7
UniSFDA$_{E\to S}$	**94.3**	61.7	85.8	35.1	2.5	38.2	**53.9**	44.5	86.4	88.8	61.4	27.1	88.8	53.0	**49.5**	55.4	**57.9**	**65.4**
*UniSFDA$_S$**	89.6	51.0	85.8	15.3	2.6	48.9	54.6	42.2	88.2	88.6	71.9	36.3	80.2	53.8	40.3	35.2	55.3	62.9
*UniSFDA$_D$**	89.1	49.3	85.7	23.8	0.9	48.8	57.2	38.0	88.1	87.7	69.2	27.6	89.0	49.6	36.2	47.0	55.5	62.6
*UniSFDA$_V$**	89.8	52.1	86.7	28.4	3.7	51.4	56.1	43.2	88.2	91.5	70.2	31.9	87.1	40.2	34.0	46.9	56.3	62.9
*UniSFDA$_{E\to S}$**	90.4	54.1	87.4	36.8	1.1	51.6	60.2	54.3	88.7	90.6	71.5	30.1	89.6	52.4	41.3	54.3	59.6	66.5

mIoU* denotes the average score across 13 categories, excluding wall, fence, and pole. UniSFDA and UniSFDA* indicate that DeepLabv2 and SegFormer are used as the source models, respectively. The best score in each column is highlighted: bold values correspond to results using the DeepLabv2 source model, while underlined values correspond to results using the SegFormer source model.

Table 3. Stage One Model Justification of Our Proposed Framework on GTA5 → Cityscapes

Model	Offline PL	Online PL	Intra-Domain FDA	mIoU	Gain
Source	-	-	-	44.0	-
SegFormer	✓			51.1	+7.1
SegFormer	✓	✓		54.9	+10.9
SegFormer	✓	✓	✓	55.4	+11.4
DFormer	✓			52.3	+8.3
DFormer	✓	✓		55.8	+11.8
DFormer	✓	✓	✓	56.6	+12.6
ViT-Adapter	✓			51.9	+7.9
ViT-Adapter	✓	✓		56.7	+12.7
ViT-Adapter	✓	✓	✓	57.8	+13.8

Source model: DeepLabv2.

is to train a source model with strong generalization ability to unseen domains. Consequently, the results in Table 5 reflect the quality of the PLs generated by the source model when different intra-domain augmentation techniques, as described in Section 3.4, are applied. As can be seen, with only geometric distortions such as random cropping and flipping, the source model achieved an mIoU of 43.0% on the target domain. Introducing photometric distortions improved the mIoU to

Table 4. Stage Two Model Justification of Our Proposed Self-Training Framework on GTA5 → Cityscapes

Stage-1 model ⟶ Stage-2 model	Confidence	Uncertainty	mIoU	Gain
SegFormer ⟶ SegFormer			56.8	-
SegFormer ⟶ SegFormer	✓		56.6	−0.2
SegFormer ⟶ SegFormer	✓	✓	58.2	+1.4
DFormer ⟶ SegFormer	✓	✓	60.9	+4.1
ViT-Adapter ⟶ SegFormer	✓	✓	60.1	+3.3
Late Fusion			57.9	+1.1
Ensembles ⟶ SegFormer	✓		61.6	+4.8
Ensembles (Intersection) ⟶ SegFormer	✓	✓	60.3	+3.5
Ensembles (Union) ⟶ SegFormer	✓	✓	59.7	+2.9
Ensembles ⟶ SegFormer	✓	✓	61.8	+5.0

Source model: DeepLabv2.

Table 5. Impact Evaluation of Intra-Domain Augmentations on the Generalization Performance of the SegFormer Source Model on GTA5 → Cityscapes

Geometric Distortion	Photometric Distortion	CutMix and ClassMix	Intra-Domain FDA	Weather and Cartoon	mIoU	Gain
✓					43.0	-
✓	✓				46.3	+3.3
✓	✓	✓			47.5	+4.5
✓	✓		✓	✓	47.3	+4.3
✓	✓	✓	✓	✓	49.0	+6.0

46.3%. By incorporating the remaining augmentations, we obtained the best source model, achieving an mIoU of 49.0% on the target domain. This experiment demonstrates that the generalizability of SegFormer can be significantly improved through simple and efficient non-learning-based intra-domain augmentations.

Comparison to Existing Domain Generalization Methods. Table 6 shows a comparison between existing domain generalization methods. Our SegFormer source model, trained with non-learning-based augmentations, is able to outperform earlier work that diversifies training samples using learned styles. However, recent efforts that focused on utilizing high-resolution images [24], transfer learning from ImageNet features [23] or foundation models such as CLIP, SAM, and LLM [4], have boosted domain-generalized semantic segmentation to the next level. Fortunately, our proposed *UniSFDA* focuses on target domain adaptation, which is flexible to start with various source models with different architectures. We plan to explore CLOUDS [4] as the source model and further advance the field of source-free domain adaptive semantic segmentation in the future.

Impact of Key Parameters in the Uncertainty-Aware PL Fusion. There are two key parameters m_c and m_u for calculating the per-class thresholds σ_k and τ_k in Equation (2). Recall that σ_k and τ_k are the thresholds for selecting hard PLs based on confidence and uncertainty scores, respectively. We set them to 33%, 66%, and 100% to retain different proportions of offline PLs and compared the results in Figure 3. This experiment is conducted in stage one self-training, using GTA5 as the source dataset and SegFormer as both the teacher and student models. We observe that the best performance was obtained with $m_c = 66\%$ and $m_u = 66\%$. When the retained proportions are

Table 6. Comparison of Domain Generalization Methods Trained on GTA and Evaluated on Cityscapes

Method	Backbone	Training Strategy	mIoU	Gain
GtA [32]	ResNet-101	Data aug., multi-head	44.0	-
AdvStyle [83]		Adversarial style aug.	44.5	+0.5
SHADE [81]		Style-hallucinated aug.	46.7	+2.7
CLOUDS [4]		CLIP, SAM, LLM	50.6	+4.2
SHADE [81]	MiT-5B	Style-hallucinated aug.	48.5	+4.5
DAFormer [23, 25]		Data aug., RCS, FD	51.6	+7.6
HRDA [24, 25]		Data aug., RCS, FD, MR	55.5	+11.5
CLOUDS [4]		CLIP, SAM, LLM	58.1	+14.1
Ours		Non-learning-based aug.	49.0	+5.0
CLOUDS [4]	ConvNext-L [39]	CLIP, SAM, LLM	60.2	+16.2

RCS, FD, and MR refer to rare class sampling, thing-class ImageNet feature distance, and multi-resolution training.

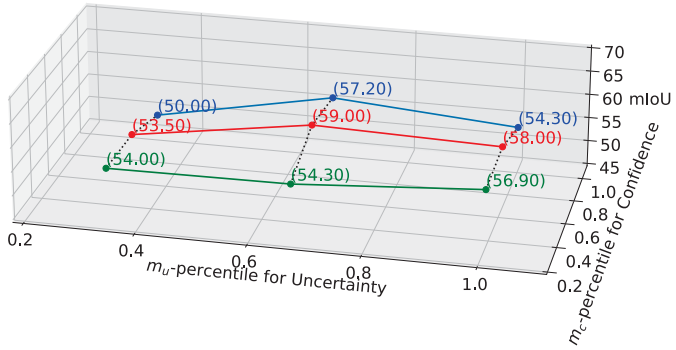

Fig. 3. Ablation study on the impact of m_c and m_u for reliable offline PL selection in self-training (source domain: GTA5, source model: SegFormer, target model: SegFormer).

too small, there are insufficient diverse training samples; conversely, when they are too large, the dataset contains more noisy PLs. Both cases will lead to decreased performance.

Impact of Reference Image Selection for Intra-Domain FDA. For our proposed intra-domain FDA, we use the entire training set as the reference image set by default, from which a random image is selected as the reference each time. However, the performance of the intra-domain FDA may depend on the choice of reference images. Table 7 presents the performance comparison of our stage-1 model trained using different reference image selections for intra-domain FDA. When no intra-domain FDA is applied, the model achieves an mIoU of 58.2%. Using the entire training set as the reference image set improves the mIoU to 59.0%. We further experimented with selecting a small and balanced subset from the training data as the reference image set, by sampling frames from each video sequence in the Cityscapes dataset. The experimental results indicate that this strategy is beneficial and boosts the mIoU to 61.2%.

Comparison of Segmentation Model Options during Inference. We compare the effectiveness and efficiency of the final-stage segmentation model options in Table 8. SegFormer is relatively light-weight, requiring only an image as input. DFormer is a multi-modal segmentation model, which requires both image and depth data as inputs. ViT-Adapter leverages a pretrained ViT model, whose

Table 7. Ablation Study on the Impact of Reference Image Selection for Intra-Domain FDA

Reference Images for Intra-Domain FDA	PL Selection	mIoU
None (without intra-domain FDA)	$m_c = 0.66$, $m_u = 0.66$	58.2
Whole training set (default)	$m_c = 0.66$, $m_u = 0.66$	59.0
1 frame from each training sequence	$m_c = 0.66$, $m_u = 0.66$	61.2
2 frames from each training sequence	$m_c = 0.66$, $m_u = 0.66$	59.8
Half of the frames from each training sequence	$m_c = 0.66$, $m_u = 0.66$	60.3

Source domain: GTA5, Source model: SegFormer, Target model: SegFormer.

Table 8. Comparison of the Effectiveness and Efficiency of the Final-Stage Segmentation Model during Inference on GTA5 → Cityscapes

Source Model	Stage-1 Model	Stage-2 Model	Input		Runtime per Image	Gflops	mIoU
			Image	Depth			
DeepLabv2	Ensembles	SegFormer	✓	✗	100 ms	799	61.8
DeepLabv2	Ensembles	DFormer	✓	✓	60 ms	447	61.5
DeepLabv2	Ensembles	ViT-Adapter	✓	✗	597 ms	4,004	61.6
SegFormer	Ensembles	SegFormer	✓	✗	100 ms	799	65.4
SegFormer	Ensembles	DFormer	✓	✓	60 ms	447	66.8
SegFormer	Ensembles	ViT-Adapter	✓	✗	597 ms	4,004	66.6

scale is considerably larger than that of SegFormer and DFormer. An interesting observation is that DFormer and ViT-Adapter did not outperform SegFormer when DeepLabv2 was used as the source model. A possible explanation is that the capabilities of these more powerful models were constrained by noise in the generated PLs. By adopting SegFormer as the source model, more accurate PLs were produced, allowing DFormer and ViT-Adapter to surpass SegFormer. However, this came at the cost of requiring extra modalities or significantly larger model size, which may not always be feasible in real-world applications due to missing modalities or limited computational resources. In terms of efficiency, we report the runtime per image and Gflops of different models. We observe that ViT-Adapter is significantly less efficient than the other two models. Although DFormer is more lightweight, it requires depth information, which may not be available in practice during inference. It is also worth noting that ViT-Adapter ($UniSFDA_V$) outperformed SegFormer ($UniSFDA_S$) by 4.6% in stage 1, as shown in Table 1. However, in stage 2, the gap between the two models was substantially reduced to 1.2%, demonstrating the effectiveness of our proposed two-stage self-training framework.

System Efficiency during Training. We further evaluate system efficiency during training in Table 9, which reports the number of trainable parameters and the actual training time of the models in our proposed multi-stage framework. In stage 1, we train the two auxiliary models, DFormer and ViT-Adapter, which require 9.9 minutes and 34.2 minutes per epoch, respectively. In stage 2, we train SegFormer using the PLs generated by DFormer and ViT-Adapter, which takes 19.3 minutes per epoch. Recall that all models are trained on four NVIDIA L40 GPUs with a mini-batch size of 4 for 60 epochs. Compared with a single-stage setup in which a single SegFormer is trained using the source model, our proposed multi-stage framework incurs an additional training cost of 44.1 hours on the Cityscapes dataset, while achieving an mIoU improvement of 3.6–6.4%.

Table 9. Trainable Parameters and Training Time Analysis of the Models in the Proposed Multi-Stage Framework

Stage	Model	Trainable Parameter	Training Time (Epoch/Total)
Stage 1	DFormer	39.0 M	9.9 minute/9.9 hour
	ViT-Adapter	570.4 M	34.2 minute/34.2 hour
Stage 2	SegFormer	84.6 M	19.3 minute/19.3 hour

Table 10. Integration of Our Method with Uni-UVPT, Demonstrating That Foundation Models Can Further Enhance Existing SFDA Approaches

Method	GTA5 → Cityscapes	SYNTHIA → Cityscapes	
	mIoU	mIoU	mIoU*
Uni-UVPT	56.9	53.8	60.4
UniSFDA (Uni-UVPT)	57.9	54.6	61.7
UniSFDA *(Uni-UVPT)	59.0	56.9	63.6

mIoU* denotes the averaged score across the 13 categories on SYNTHIA → Cityscapes.

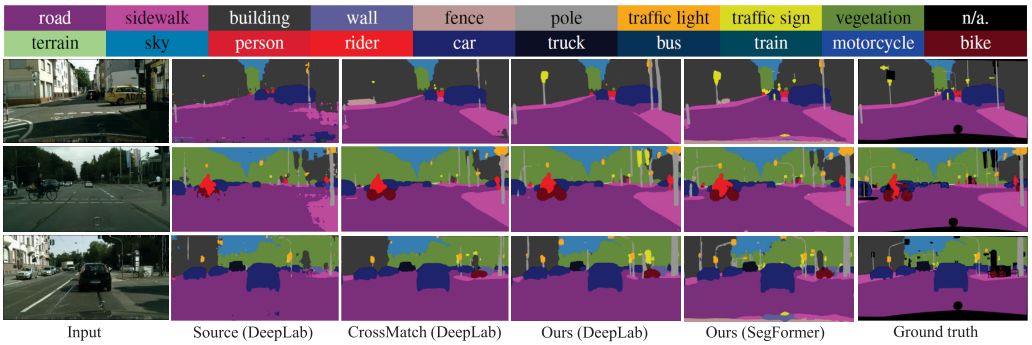

Fig. 4. Qualitative results of source-free semantic segmentation on the Cityscapes dataset. From left to right: input image, output from the DeepLabv2 source model, output from the CrossMatch model [73] trained with the DeepLabv2 source model, outputs from our proposed model trained with the DeepLabv2 and SegFormer source models, respectively, and the ground-truth segmentation mask.

Integration with Existing SFDA Approaches. While existing SFDA methods mostly focus on PL denoising techniques and do not leverage foundation models, we demonstrate that foundation models can further enhance existing SFDA approaches by integrating our method with a recent approach, Uni-UVPT, using Swin-B as the backbone. As shown in Table 10, the original Uni-UVPT achieved a mIoU of 56.9% on GTA5 → Cityscapes and 53.8% on SYNTHIA → Cityscapes. By using our stage-1 models, DFormer and ViT-Adapter, to generate PLs via the proposed PL aggregation method in Equation (6), in conjunction with the PL correction strategy of Uni-UVPT, we improve the mIoU to 57.9% and 54.6% when using DeepLabv2 as the source model, and further to 59.0% and 56.9% when using SegFormer as the source model. The results indicate that our two-stage self-training framework can be seamlessly integrated with existing PL denoising approaches to achieve additional performance gains.

Visualizations. Finally, we conducted a qualitative evaluation of our method. As illustrated in Figure 4, we visualized the prediction results obtained by the DeepLabv2 source model, the CrossMatch model trained with DeepLabv2 source model, and our method trained with both

DeepLabv2 and SegFormer source models. As observed, the predictions generated by the source model can be noisy due to the domain gap between the source and target dataset, particularly in source-free settings, where the source dataset is unavailable during training. The CrossMatch method significantly reduces label noise by leveraging depth information as an additional input. However, it still struggles to capture fine details accurately. In comparison, our approach achieves more precise class-wise predictions, particularly for small objects such as poles, traffic lights, and traffic signs.

5 Conclusion

This article proposes a unified self-training framework for robust source-free domain adaptive semantic segmentation. We identify three underexplored aspects in prior work, namely uncertainty-aware PL fusion, transfer learning-based adaptation, and intra-domain style augmentation, and investigate them extensively to collectively enhance segmentation performance and test-time efficiency on the target domain. Our approach achieves new state-of-the-art results in urban scene understanding, obtaining an mIoU of 61.8% and 57.9% on the Cityscapes dataset when adapting from GTA5 or SYNTHIA benchmarks using a DeepLabv2 source model. The proposed two-stage self-training framework not only produces more reliable PLs but also ensures the feasibility and efficiency of the test-time segmentation model. It leverages recent advances in cross-modal learning and large foundation models through transfer learning, and can be seamlessly integrated with a wide range of models and self-training methodologies.

References

[1] Ibrahim Batuhan Akkaya and Ugur Halici. 2022. Self-training via metric learning for source-free domain adaptation of semantic segmentation. arXiv:2212.04227. Retrieved from https://arxiv.org/abs/2212.04227

[2] Nikita Araslanov and Stefan Roth. 2021. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, 15384–15394.

[3] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. 2020. Source-relaxed domain adaptation for image segmentation. In *MICCAI*, 490–499.

[4] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. 2024. Collaborating foundation models for domain generalized semantic segmentation. In *CVPR*, 3108–3119.

[5] Yihong Cao, Hui Zhang, Xiao Lu, Zheng Xiao, Kailun Yang, and Yaonan Wang. 2023. Towards source-free domain adaptive semantic segmentation via importance-aware and prototype-contrast learning. arXiv:2306.01598. Retrieved from https://arxiv.org/abs/2306.01598

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2017), 834–848.

[7] Mu Chen, Minghan Chen, and Yi Yang. 2024. UAHOI: Uncertainty-aware robust interaction learning for HOI detection. *Computer Vision and Image Understanding* 247 (2024), 104091.

[8] Mu Chen, Zhedong Zheng, and Yi Yang. 2024. Transferring to real-world layouts: A depth-aware framework for scene adaptation. In *ACM Multimedia*, 399–408.

[9] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. 2023. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *ACM Multimedia*, 1905–1914.

[10] Yizhen Chen and Haifeng Hu. 2021. Y-Net: Dual-branch joint network for semantic segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 4 (2021), 1–22.

[11] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. 2017. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 1992–2001.

[12] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. 2023. Vision transformer adapter for dense predictions. In *ICLR*.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 3213–3223.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[15] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. 2019. SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation. In *ICCV*, 982–991.

[16] Yan Fang, Feng Zhu, Bowen Cheng, Luoqi Liu, Yao Zhao, and Yunchao Wei. 2023. Locating noise is halfway denoising for semi-supervised segmentation. In *ICCV*, 16612–16622.

[17] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*. PMLR, 1050–1059.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.

[19] Adam Goodge, Wee Siong Ng, Bryan Hooi, and See Kiong Ng. 2025. Spatio-temporal foundation models: Vision, challenges, and opportunities. arXiv:2501.09045. Retrieved from https://arxiv.org/abs/2501.09045

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*. PMLR, 1321–1330.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 1989–1998.

[23] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 9924–9935.

[24] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*.

[25] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2023. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. In *IEEE TPAMI*.

[26] Wenmiao Hu, Yifang Yin, Ying Kiat Tan, An Tran, Hannes Kruppa, and Roger Zimmermann. 2024. GAN-assisted road segmentation from satellite imagery. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 1 (2024), 1–29.

[27] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. 2021. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Advances in Neural Information Processing System*, Vol. 34, 3635–3649.

[28] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. 2023. Style projected clustering for domain generalized semantic segmentation. In *CVPR*, 3061–3071.

[29] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, et al. 2020. imgaug. Retrieved February 1, 2020 from https://github.com/aleju/imgaug

[30] Patrick Kage, Jay C. Rothenberger, Pavlos Andreadis, and Dimitrios I. Diochnos. 2024. A review of pseudo-labeling for computer vision. arXiv:2408.07221. Retrieved from https://arxiv.org/abs/2408.07221

[31] Sunghwan Kim, Dae-Hwan Kim, and Hoseong Kim. 2023. Texture learning domain randomization for domain generalized segmentation. In *ICCV*, 677–687.

[32] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R. Venkatesh Babu. 2021. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, 7046–7056.

[33] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. 2022. WildNet: Learning domain generalized semantic segmentation from the wild. In *CVPR*, 9936–9946.

[34] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. 2020. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, 440–456.

[35] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 6936–6945.

[36] Feng Lin, Bin Li, Wengang Zhou, Houqiang Li, and Yan Lu. 2020. Single-stage instance segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 3 (2020), 1–19.

[37] Wenxi Liu, Jiaxin Cai, Qi Li, Chenyang Liao, Jingjing Cao, Shengfeng He, and Yuanlong Yu. 2024. Learning nighttime semantic segmentation the hard way. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 7 (2024), 1–23.

[38] Yuang Liu, Wei Zhang, and Jun Wang. 2021. Source-free domain adaptation for semantic segmentation. In *CVPR*, 1215–1224.

[39] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *CVPR*, 11976–11986.

[40] Zhihe Lu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. 2023. Uncertainty-aware source-free domain adaptive semantic segmentation. *IEEE Transactions on Image Processing* 32 (2023), 4664–4676.
[41] Xin Luo, Wei Chen, Zhengfa Liang, Longqi Yang, Siwei Wang, and Chen Li. 2023. Crots: Cross-domain teacher–student learning for source-free domain adaptive semantic segmentation. *International Journal of Computer Vision* 132 (2023), 20–39.
[42] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *ICCV*, 2507–2516.
[43] Xinhong Ma, Yiming Wang, Hao Liu, Tianyu Guo, and Yunhe Wang. 2023. When visual prompt tuning meets source-free domain adaptive semantic segmentation. In *Advances in Neural Information Processing System*, Vol. 36, 6690–6702.
[44] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. 2020. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*.
[45] Luke Melas-Kyriazi and Arjun K. Manrai. 2021. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *CVPR*, 12435–12445.
[46] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 1369–1378.
[47] An-Tao Pan, Yawei Luo, Yi Yang, and Jun Xiao. 2022. DUDA: Online-offline dual domain adaption for semantic segmentation. In *BMVC*, 585.
[48] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. 2021. AUGCO: Augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. arXiv:2107.10140. Retrieved from https://arxiv.org/abs/2107.10140
[49] S. Prabhu Teja and François Fleuret. 2021. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, 9613–9623.
[50] Xiaofeng Qu, Li Liu, Lei Zhu, Liqiang Nie, and Huaxiang Zhang. 2024. Instance-level adversarial source-free domain adaptive person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 7 (2024), 1–22.
[51] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *ECCV*, 102–118.
[52] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. 2021. Defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*. Retrieved from https://openreview.net/forum?id=-ODN6SbiUU
[53] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 3234–3243.
[54] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. 2018. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, 687–704.
[55] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. 2020. Two-phase PL densification for self-training based domain adaptation. In *ECCV*, 532–548.
[56] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing System*, Vol. 30, 1195–1204.
[57] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 7472–7481.
[58] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2517–2526.
[59] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 7364–7373.
[60] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. 2021. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, 8515–8525.
[61] Xizhong Wang, Rui Liu, Xin Yang, Qiang Zhang, and Dongsheng Zhou. 2024. MCFNet: Multi-attentional class feature augmentation network for real-time scene parsing. *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, 6 (2024), 1–17.
[62] Yan Wang, Hong Xie, Jinyang He, Xiaoyu Shi, and Mingsheng Shang. 2025. Cross-domain semantic transfer for domain generalization. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 5 (2025), 1–24.
[63] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. 2021. The temporal opportunist: Self-supervised multi-frame monocular depth. In *CVPR*, 1164–1174.
[64] Hongchen Wei and Zhenzhong Chen. 2025. Improving domain generalization for image captioning with unsupervised prompt learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 21 (2025), 1–23.

[65] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. 2024. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *CVPR*, 28619–28630.
[66] Chao Wen, Chen Wei, Yuhua Qian, Xiaodan Song, and Xuemei Xie. 2025. Prompt-based invertible mapping alignment for unsupervised domain adaptation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21, 5 (2025), 1–17.
[67] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing System*, Vol. 34, 12077–12090.
[68] Cheng-Yu Yang, Yuan-Jhe Kuo, and Chiou-Ting Hsu. 2022. Source free domain adaptation for semantic segmentation via distribution transfer and adaptive class-balanced self-training. In *ICME*, 1–6.
[69] Liwei Yang, Xiang Gu, and Jian Sun. 2023. Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, 10789–10797.
[70] Yanchao Yang and Stefano Soatto. 2020. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 4085–4095.
[71] Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. 2021. Source data-free unsupervised domain adaptation for semantic segmentation. In *ACM International Conference on Multimedia*, 2233–2242.
[72] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. 2024. DFormer: Rethinking RGBD representation learning for semantic segmentation. In *ICLR*.
[73] Yifang Yin, Wenmiao Hu, Zhenguang Liu, Guanfeng Wang, Shili Xiang, and Roger Zimmermann. 2023. CrossMatch: Source-free domain adaptive semantic segmentation via cross-modal consistency training. In *ICCV*, 21786–21796.
[74] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. 2021. Domain adaptive semantic segmentation without source data. In *ACM International Conference on Multimedia*, 3293–3302.
[75] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 6023–6032.
[76] Qi Zang, Shuang Wang, Dong Zhao, Yang Hu, Dou Quan, Jinlong Li, Nicu Sebe, and Zhun Zhong. 2024. Generalized source-free domain-adaptive segmentation via reliable knowledge propagation. In *ACM Multimedia*, 5967–5976.
[77] Kai Zhang, Yifan Sun, Rui Wang, Haichang Li, and Xiaohui Hu. 2021. Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. arXiv:2112.00295. Retrieved from https://arxiv.org/abs/2112.00295
[78] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 2021. Prototypical PL denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 12414–12424.
[79] Dong Zhao, Shuang Wang, Qi Zang, Licheng Jiao, Nicu Sebe, and Zhun Zhong. 2024. Stable neighbor denoising for source-free domain adaptive segmentation. In *CVPR*, 23416–23427.
[80] Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, and Licheng Jiao. 2023. Towards better stability and adaptability: Improve online self-training for model adaptation in semantic segmentation. In *CVPR*, 11733–11743.
[81] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*, 535–552.
[82] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. 2024. Style-hallucinated dual consistency learning: A unified framework for visual domain generalization. *International Journal of Computer Vision* 132, 3 (2024), 837–853.
[83] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. In *Advances in Neural Information Processing System*, Vol. 35, 338–350.

Received 14 September 2025; revised 28 December 2025; accepted 30 January 2026