Semantics, Distortion, and Style Matter: Towards Source-free UDA for Panoramic Segmentation

Xu Zheng[1] Pengyuan Zhou[3] Athanasios V. Vasilakos[4] Lin Wang[1,2*]
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST [3] Aarhus University [4] University of Agder
zhengxu128@gmail.com, pengyuan.zhou@ece.au.dk, th.vasilakos@gmail.com, linwang@ust.hk
Project Page: https://vlislab22.github.io/360SFUDA/

Abstract

*This paper addresses an interesting yet challenging problem– source-free unsupervised domain adaptation (SFUDA) for pinhole-to-panoramic semantic segmentation–given only a pinhole image-trained model (*i.e.*, source) and unlabeled panoramic images (*i.e.*, target). Tackling this problem is nontrivial due to the semantic mismatches, style discrepancies, and inevitable distortion of panoramic images. To this end, we propose a novel method that utilizes Tangent Projection (TP) as it has less distortion and meanwhile slits the equirectangular projection (ERP) with a fixed FoV to mimic the pinhole images. Both projections are shown effective in extracting knowledge from the source model. However, the distinct projection discrepancies between source and target domains impede the direct knowledge transfer; thus, we propose a panoramic prototype adaptation module (PPAM) to integrate panoramic prototypes from the extracted knowledge for adaptation. We then impose the loss constraints on both predictions and prototypes and propose a cross-dual attention module (CDAM) at the feature level to better align the spatial and channel characteristics across the domains and projections. Both knowledge extraction and transfer processes are synchronously updated to reach the best performance. Extensive experiments on the synthetic and real-world benchmarks, including outdoor and indoor scenarios, demonstrate that our method achieves significantly better performance than prior SFUDA methods for pinhole-to-panoramic adaptation.*

1. Introduction

The comprehensive scene perception abilities of 360° cameras have made them highly popular for applications, such as autonomous driving [1]. In contrast to pinhole cameras that capture 2D planer images with a limited field-of-view (FoV), 360° cameras offer a much wider FoV of 360° × 180°. As a result, research on panoramic semantic segmentation [42, 43, 46, 48, 49] has been actively explored to achieve dense scene understanding for intelligent systems.

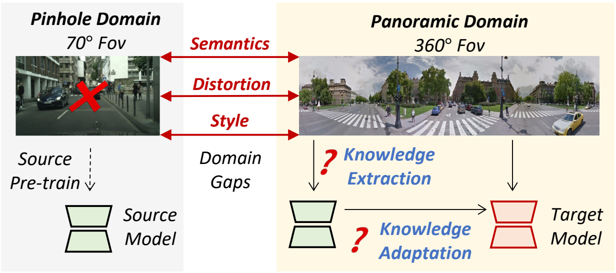

Figure 1. We address a new problem of achieving source-free pinhole-to-panoramic adaptation for segmentation.

Generally, the spherical data captured by the 360° cameras is always projected into 2D planar representations, *e.g.*, Equirectangular Projection (ERP), to be aligned with the existing imaging pipeline [1] while preserving the omnidirectional information [1]. However, ERP suffers from the inevitable distortion and object deformation due to the non-uniformly distributed pixels [59]. Meanwhile, learning effective panoramic segmentation models is often impeded by the lack of large precisely labeled datasets due to the difficulty of annotation. For these reasons, some unsupervised domain adaptation (UDA) methods [49, 50, 59] have been proposed to transfer the knowledge from the pinhole image domain to the panoramic image domain. In some crucial application scenarios, *e.g.*, autonomous driving, source datasets are not always accessible due to privacy and commercial issues, such as data portability and transmission costs. One typical example is the recent large model, SAM [19], which brings significant progress in instance segmentation for pinhole images; however, the source datasets are too large (10TB) to be reused in end-tasks, such as [20].

[*] Corresponding author.

[1] In this paper, omnidirectional and panoramic images are interchangeably used, and ERP images often indicate panoramic images.

Motivation: In this paper, we probe an interesting yet challenging problem: *source-free UDA (SFUDA) for panoramic segmentation, in which only the source model (pretrained with pinhole images) and unlabeled panoramic images are available*. As shown in Fig. 1 (a), different from existing SFUDA methods, *e.g*., [25, 41, 44] for the pinhole-to-pinhole image adaptation, transferring knowledge from the pinhole-to-panoramic image domain is hampered by: **1)** semantic mismatch caused by the different FoV between the pinhole and 360° cameras, *i.e*., 70° vs. 360°; **2)** inevitable distortion of the ERP; **3)** style discrepancies caused by the distinct camera sensors and captured scenes. In Tab. 2, we show that naively adapting existing SFUDA methods to our problem leads to a limited performance boost.

Contributions: To this end, we propose a novel SFUDA method that effectively extracts knowledge from the source model with only panoramic images and transfers the knowledge to the target panoramic domain. *Our key idea is to leverage the multi-projection versatility of 360° data for efficient domain knowledge transfer*. Our method enjoys two key technical contributions. Specifically, we use Tangent Projection (TP) and divide the ERP images into patches with a fixed FoV, dubbed Fixed FoV Projection (FFP), to extract knowledge from the source model with less distortion and similar FoV to the pinhole images. Both projections make it possible to effectively extract knowledge from the source model. However, directly transferring the extracted knowledge to the target model is hardly approachable due to the distinct projection gaps. Thus, we propose a panoramic prototype adaptation module (PPAM) to obtain *class-wise semantic prototypes* from the features and predictions of the source model with TP and FFP images (Sec. 3.2). Then, these prototypes are integrated together to obtain the global panoramic prototypes for knowledge adaptation, which is updated across the adaptation procedure. Moreover, our proposed PPAM also fine-tunes the source model to promote better knowledge extraction using prototypes extracted from FFP images. Aligning the prototypes from each FFP image enables the source model to become more aware of distortion and semantics across the FoV.

We initially apply both prediction-level and prototype-level loss constraints to facilitate knowledge transfer to the unlabeled target panoramic domain. Concretely, the FFP predictions of the source model are rebuilt together to provide a pseudo-supervision signal for the target model. The prototype-level loss constraint is performed between the panoramic prototypes from PPAM and the prototypes from the target model's features and predictions on the ERP images. Moreover, knowledge from the source model is not limited to predictions and prototypes, high-level features also contain crucial image characteristics that can enhance the performance of the target model. Consequently, we propose a Cross-Dual Attention Module (**CDAM**) that *aligns spatial and channel characteristics between domains* to fully utilize the knowledge from the source model and address the style discrepancy problem (Sec. 3.3). Specifically, CDAM reconstructs the source model features from FFP images to provide a panoramic perception of the surrounding environment and aligns them with the ERP features from the target model for effective knowledge transfer.

We conduct extensive experiments on both synthetic and real-world benchmarks, including outdoor and indoor scenarios. As no directly comparable works exist, we adapt the state-of-the-art (SoTA) SFUDA methods [14, 18, 21, 25, 41, 51] – designed for pinhole-to-pinhole image adaptation – to our problem in addressing the panoramic semantic segmentation. The results show that our framework significantly outperforms these methods by large margins of +6.37%, +11.47%, and +10.99% on three benchmarks. We also evaluate our method against UDA methods [49, 50, 58, 59], using the source pinhole image, the results demonstrate its comparable performance.

2. Related Work

2.1. Source-free UDA for Segmentation

UDA aims to mitigate the impact of domain shift caused by data distribution discrepancies in downstream computer vision tasks, such as semantic segmentation [2, 6–9, 13, 17, 30, 32, 33, 36, 37, 40, 52, 55–57, 60, 61]. However, the source domain data may not always be accessible due to the privacy protection and data storage concerns. Intuitively, source-free UDA (SFUDA) [18, 21, 45] methods are proposed to adapt source models to a target domain without access to the source data. Existing SFUDA methods for semantic segmentation primarily focus on source data estimation [41, 44] or self-training [4, 21, 25, 54] for pinhole images. *In this paper, we make the first attempt at achieving SFUDA from the pinhole image domain to the panoramic domain*. This task is nontrivial to be tackled due to the semantic mismatches, style discrepancies, and inevitable distortion of panoramic images. Unlike these methods that focus on the source domain data estimation [25, 44], we propose a novel SFUDA method that effectively extracts knowledge from the source model with only panoramic images and transfers the knowledge to the target panoramic image domain. Experiments also show that naively applying these methods leads to less optimal performance (See Tab. 2).

2.2. UDA for Panoramic Semantic Segmentation

It can be classified into three types, including adversarial training [10, 16, 31, 34, 59], pseudo labeling [24, 38, 47, 53] and prototypical adaptation methods [49, 50]. Specifically, the first line of research applies alignment approaches to capture the domain invariant characteristics of images [16, 22, 29], feature [5, 15, 16, 59] and predictions [26, 28].

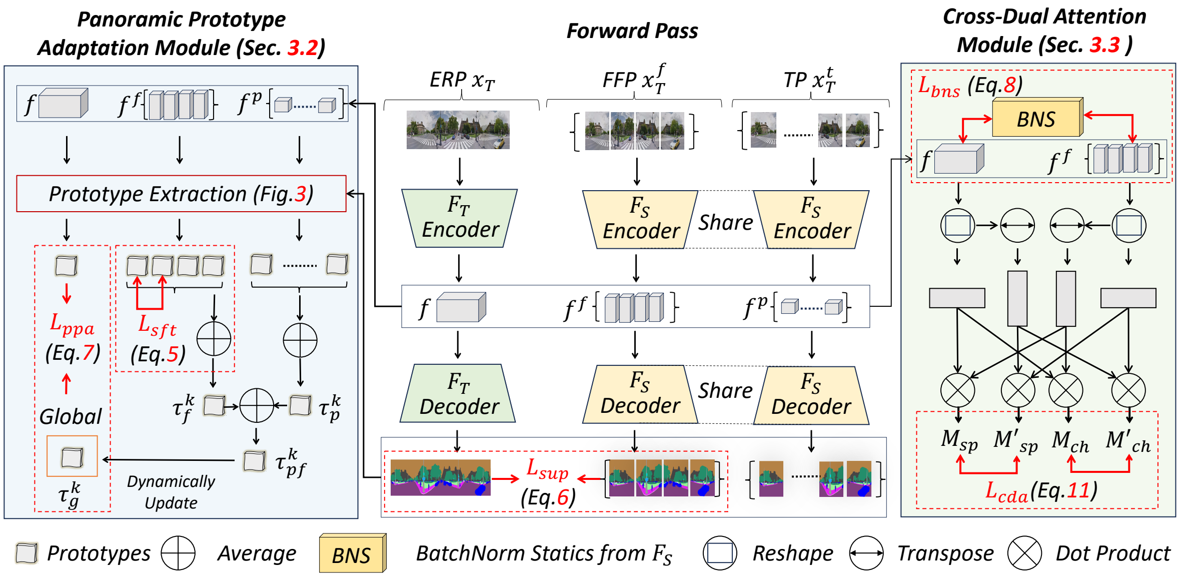

Figure 2. Overall framework of our proposed SFUDA for panoramic semantic segmentation.

The second type of methods generates pseudo labels for the target domain training. The last line of research, *e.g.*, Mutual Prototype Adaption (MPA) [49], mutually aligns the high-level features with the prototypes between domain. However, these methods treat panoramic images as pinhole images when extracting prototypes, ignoring the intricate semantic, object correspondence, and distortion information brought by the panoramic FoV. *We are the first to address the SFUDA problem for panoramic segmentation. Considering the distinct projection discrepancies between source and target domains, we propose a PPAM to integrate the global panoramic prototypes from the extracted knowledge for adaptation.*

3. Methodology

3.1. Overview

The overall framework for panoramic segmentation is shown in Fig. 2. With only the source model F_S available and given the unlabeled panoramic image data D_T, we aim to train a target model F_T that adapts knowledge from F_S to the common K categories across both domains.

Unlike the pinhole image-to-image adaptation [25, 41, 44], pinhole-to-panoramic image domain adaptation is hampered by three key factors, specifically: semantic mismatch due to FoV variations (70° vs. 360°), inevitable distortion in ERP, and ubiquitous style discrepancies in unsupervised domain adaptation (UDA) (refer to Fig.1 (a)). Therefore, naively applying existing SFUDA methods exhibits suboptimal segmentation performance (See Tab. 2), while UDA methods with source data, *e.g.*, [25] for panoramic segmentation do not account for the semantic mismatch between the pinhole and panoramic images. Intuitively, the key challenges are : **1)** how to extract knowledge from the source model with only panoramic images and **2)** how to transfer knowledge to the target panoramic image domain.

Our key idea *is to leverage the multi-projection versatility of 360° data for efficient domain knowledge transfer.*

Concretely, to address the first challenge (Sec. 3.2), we use the Tangent Projection (TP) which is characterized by a reduced distortion issue compared to the ERP images [12] to extract knowledge from the source model. Concurrently, ERP images are segmented into discrete patches, each possessing a constant FoV to mimic the pinhole images, dubbed Fixed FoV Projection (FFP). Both projections make it possible to effectively extract knowledge from the source model. The distinct projection formats make it impossible to directly transfer knowledge between domains, thus we propose a Panoramic Prototype Adaptation Module (PPAM) to obtain panoramic prototypes for adaptation. To address the second challenge (Sec. 3.3), we first impose prediction and prototype level loss constraints, and propose a Cross-Dual Attention Module (CDAM) at the feature level to transfer knowledge and further address the style discrepancies.

3.2. Knowledge Extraction

As depicted in Fig. 2, given the target domain (*i.e.*, panoramic domain) ERP images $D_T = \{x_T | x_T \in \mathbf{R}^{H\times W\times 3}\}$, we first project them into TP images $D_T^t = \{x_T^t | x_T^t \in \mathbf{R}^{h\times w\times 3}\}$ and FFP images $D_T^f = \{x_T^f | x_T^f \in \mathbf{R}^{H\times W/4\times 3}\}$ for effectively extracting knowledge from the source model. Note that one ERP image corresponds to 18

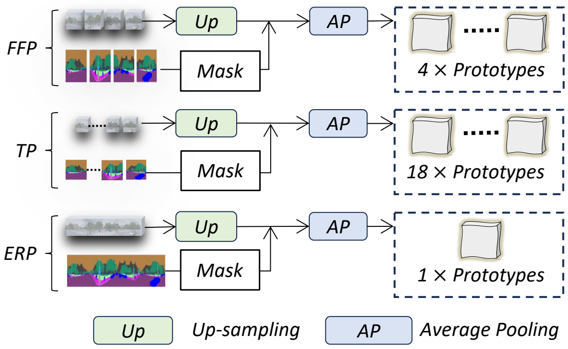

Figure 3. Illustration of the prototype extraction (PE) in the panoramic prototype adaptation module (PPAM).

TP images as [23, 59] and 4 FFP images with a fixed FoV of 90° (See Sec. 5). To obtain the features and predictions from the source model for knowledge adaptation, the two types of projected images are first fed into the source model with batch sampling:

$$P^p, f^p = F_S(x_T^t), \qquad P^f, f^f = F_S(x_T^f), \tag{1}$$

where f^p, f^f, P^p, and P^f are the source model features and predictions of the input TP and FFP images, respectively. For the target panoramic images, x_T is fed into F_T to obtain the target model features f and predictions P of the input batch of ERP images as $P, f = F_T(x_T)$. However, the distinct projection formats of the input data in the source and target models make it difficult to align their features directly, thus we propose a Panoramic Prototype Adaptation Module (PPAM) to obtain panoramic prototypes for adaptation.

Panoramic Prototype Adaptation Module (PPAM) Compared to prior UDA methods using prototypical adaptation, *e.g.*, MPA [49, 50], our PPAM possesses three distinct characteristics: **(a)** class-wise prototypes are obtained from TP and FFP images to alleviate distortion and semantic mismatch problems; **(b)** global prototypes are iteratively updated with prototypes from two projections during the whole training procedure; **(c)** hard pseudo-labels are softened in the high-level feature space to obtain prototypes with different projection of panoramic images, indicating that the knowledge from the source model is fully utilized.

Specifically, we project the source model predictions P^p, P^f into pseudo labels:

$$\begin{aligned} \hat{y}^p_{(h,w,k)} &= 1_{k \doteq argmax(P^p_{h,w,:})}, \\ \hat{y}^f_{(H,W/4,k)} &= 1_{k \doteq argmax(P^f_{H,W/4,:})}. \end{aligned} \tag{2}$$

Here, k denotes the semantic category. Subsequently, we obtain the class-specific masked features by integrating the up-sampled features with the corresponding pseudo labels $\hat{y}^p_{(h,w,k)}$ and $\hat{y}^f_{(H,W/4,k)}$. Notably, the prototypes $\sum_{a=1}^{18}(\tau_p^k)_a$ and $\sum_{b=1}^{4}(\tau_f^k)_b$ for TP and FFP images are obtained by masked average pooling (MAP) operation, as shown in Fig. 3. Within each projection, PPAM first integrates the prototypes:

$$\tau_p^k = avg(\sum_{a=1}^{18}(\tau_p^k)_a), \qquad \tau_f^k = avg(\sum_{b=1}^{4}(\tau_f^k)_b). \tag{3}$$

As shown in Fig. 2, τ_p^k and τ_f^k are integrated together as τ_{pf}^k to preserve the less distortion characteristics of τ_p^k and the similar scale semantics of τ_f^k. The τ_{pf}^k is then used to update the panoramic global prototype τ_g^k, which is iteratively updated with τ_{pf}^k. To obtain more accurate and reliable prototypes, we update τ_g^k and τ_{pf}^k as follows:

$$\tau_g^i = \frac{1}{i}(\tau_{pf}^k)^i + (1 - \frac{1}{i})(\tau_g^k)^{i-1}, \tag{4}$$

where $(\tau_g^k)^i$ and $(\tau_{pf}^k)^i$ are the prototypes for category k in the $i-$th training epoch, $(\tau_g^k)^{i-1}$ is the panoramic global prototype saved in the last training epoch, i is the current epoch number. The panoramic global prototype τ_g^k is then used to give supervision for the target prototype τ_t^k obtained from P and f with the same operations.

Besides extracting prototype knowledge from the source model, PPAM also fine-tunes the source model to improve the effectiveness of knowledge extraction. Specifically, since each ERP image can be projected to 4 FFP images, the source model's extracted features f_f have 4 pieces of FFP features. As the content of all the features is within the same ERP image, we propose to align the class-wise prototypes from each piece of the features in PPAM to enhance the model's performance. Concretely, the prototypes $\sum_{\alpha=1}^{4}\tau_\alpha$ of the four FFP features are obtained through the same operations with τ_g^t. Each FFP image captures a non-overlapping 90° FoV, resulting in distinct distortions, and similar content in each FFP image. Aligning the prototypes from each FFP image enhances distortion-awareness ability in the source model and helps to explore complementary semantic content in each FFP image. The MSE loss is imposed between each two of the prototypes as follows:

$$\mathcal{L}_{sft} = \sum_{\alpha \neq \beta}^{4}\{\frac{1}{K}\sum_{k \in K}((\tau_f^k)_\alpha - (\tau_f^k)_\beta)^2\}. \tag{5}$$

Note that $\mathcal{L}_{sft}$ is used to fine-tune the source model F_S.

3.3. Knowledge Adaptation

To adapt knowledge to the target domain, we impose the loss constraints on both predictions and prototypes and propose a cross-dual attention module (**CDAM**) at the feature level to better align the spatial and channel characteristics across the domains and projections. Specifically, the predictions of the FFP patch images are stitched to reconstruct an ERP image.

Method	SF	mIoU	Road	S.W.	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veget.	Terr.	Sky	Pers.	Car	Δ
PVT [39] SSL	✗	38.74	55.39	36.87	80.84	19.72	15.18	8.04	5.39	2.17	72.91	32.01	90.81	26.76	57.40	-
PVT [39] MPA	✗	40.90	70.78	42.47	82.13	22.79	10.74	13.54	1.27	0.30	71.15	33.03	89.69	29.07	64.73	-
Source w/ seg-b1	✓	35.81	63.36	24.09	80.13	**15.68**	13.39	16.26	7.42	0.09	62.45	20.20	86.05	23.02	53.37	-
SFDA w/ seg-b1 [25]	✓	38.21	68.78	30.71	80.37	5.26	18.95	20.90	5.25	2.36	70.19	23.30	$\underline{90.20}$	22.55	57.90	+2.40
ProDA w/ seg-b1 [51]	✓	37.37	68.93	30.88	80.07	4.17	18.60	19.72	1.77	1.56	70.05	22.73	**90.60**	19.71	57.04	+2.73
GTA w/ seg-b1 [21]	✓	36.00	64.61	20.04	79.04	8.06	15.36	19.86	6.02	2.13	65.77	17.75	84.56	26.71	58.13	+0.19
HCL w/ seg-b1 [18]	✓	38.38	68.82	30.41	80.37	5.88	20.18	20.10	4.23	2.11	70.50	24.74	89.89	22.65	59.04	+2.57
DATC w/ seg-b1 [41]	✓	38.54	69.48	26.96	80.68	11.64	15.24	20.10	**9.33**	0.55	66.11	24.31	85.16	30.90	60.58	+2.73
Simt w/ seg-b1 [14]	✓	37.94	68.47	29.51	79.62	6.78	19.20	19.48	2.31	1.33	68.85	$\underline{26.55}$	89.30	22.35	59.49	+2.13
Ours w/ seg-b1	✓	$\underline{41.78}$	**70.17**	**33.24**	**81.66**	$\underline{13.06}$	$\underline{23.40}$	$\underline{23.37}$	7.63	$\underline{3.59}$	$\underline{71.04}$	25.46	89.33	$\underline{36.60}$	$\underline{64.60}$	$\underline{+5.97}$
Ours w/ seg-b2	✓	**42.18**	$\underline{69.99}$	$\underline{32.28}$	$\underline{81.34}$	10.62	**24.35**	**24.29**	$\underline{9.19}$	**3.63**	**71.28**	**30.04**	88.75	**37.49**	**65.05**	**+6.37**

Table 1. Experimental results on the S-to-D scenario, the overlapped 13 classes of two datasets are used to test the UDA performance. The **bold** and underline denote the best and the second-best performance in source-free UDA methods, respectively.

The ERP image is then passed to the source model F_S to predict a pseudo label, which serves as the supervision for the ERP predictions of the target model F_T. For simplicity, we use the Cross-Entropy (CE) loss, which is formulated as:

$$\mathcal{L}_{sup} = CE(P, 1_{k \doteq argmax(\{Rebuild(P^f_{H,W/4,:})\})}). \tag{6}$$

And the prototype-level knowledge transfer loss is achieved by Mean Squared Error (MSE) loss between the panoramic global prototype τ_g^k and the target prototype τ_t^k :

$$\mathcal{L}_{ppa} = \frac{1}{K} \sum_{k \in K} (\tau_g^k - \tau_t^k)^2. \tag{7}$$

With loss $\mathcal{L}_{ppa}$, the prototypes are pushed together to transfer the source-extracted knowledge to the target domain. In summary, with the proposed PPAM, we effectively address the distortion and semantic mismatch problems at the prediction and prototype level, we now tackle the style discrepancy problem at the feature level.

Cross Dual Attention Module (CDAM). Inspired by the dual attention, focusing on spatial and channel characteristics [25], our CDAM imitates the spatial and channel-wise distributions of features to alleviate the style discrepancies. Different from [25] suggesting to minimize the distribution distance of the dual attention maps between the fake source (FFP images) and target data (ERP images), our CDAM focuses on *aligning the distribution between FFP and ERP of the panoramic images* rather than introducing additional parameters and computation cost in estimating source data. As shown in Fig. 2, we reconstruct the FFP features F^f to ensure that the rebuilt feature F' has the same spatial size as F. Before the cross dual attention operation, we apply a Batch Normalization Statics (BNS) guided constraint on F and F'. Since the BNS of the source model should satisfy the feature distribution of the source data, we align F and F' with BNS to alleviate the domain gaps as follows:

$$\begin{aligned} \mathcal{L}_{bns} =& ||\mu(F) - \bar{\mu}||_2^2 + ||\sigma^2(F) - \bar{\sigma}^2||_2^2 \\ &+ ||\mu(F') - \bar{\mu}||_2^2 + ||\sigma^2(F') - \bar{\sigma}^2||_2^2, \end{aligned} \tag{8}$$

where $\bar{\mu}$ and $\bar{\sigma}^2$ are the mean and variance parameters of the last BN layer in the source model S.

As shown in Fig. 2 (a), after aligned with BNS, the ERP feature f and the rebuilt feature f' are first reshaped to be $f \in \mathbb{R}^{N \times C}$ and $f' \in \mathbb{R}^{N \times C}$, where N is the number of pixels and C is the channel number. Then we calculate the spatial-wise attention maps $M_{sp} \in \mathbb{R}^{N \times C}$ and $M'_{sp} \in \mathbb{R}^{N \times C}$ for f and f' by:

$$\begin{aligned} \{M_{sp}\}_{ji} &= \frac{exp(f'_{[i:]} \cdot f^T_{[:j]})}{\sum_i^N exp(f'_{[i:]} \cdot f^T_{[:j]})}, \\ \{M'_{sp}\}_{ji} &= \frac{exp(f_{[i:]} \cdot f'^T_{[:j]})}{\sum_i^N exp(f_{[i:]} \cdot f'^T_{[:j]})}, \end{aligned} \tag{9}$$

where f^T is the transpose of f and $\{M\}_{ij}$ measures the impact of the i-th position on the j-th position. Similarly, the channel-wise attention maps $M_{ch} \in \mathbb{R}^{C \times C}$ and $M'_{ch} \in \mathbb{R}^{C \times C}$ can be obtained through:

$$\begin{aligned} \{M_{ch}\}_{ji} &= \frac{exp(f'^T_{[i:]} \cdot f_{[:j]})}{\sum_i^C exp(f'_{[i:]} \cdot f^T_{[:j]})}, \\ \{M'_{ch}\}_{ji} &= \frac{exp(f^T_{[i:]} \cdot f'_{[:j]})}{\sum_i^C exp(f^T_{[i:]} \cdot f'_{[:j]})}. \end{aligned} \tag{10}$$

After obtaining the spatial and channel attention maps, the CDAM loss can be calculated with the Kullback-Liibler divergence (KL divergence) as follows:

$$\mathcal{L}_{cda} = KL(M_{sp}, M'_{sp}) + KL(M_{ch}, M'_{ch}) \tag{11}$$

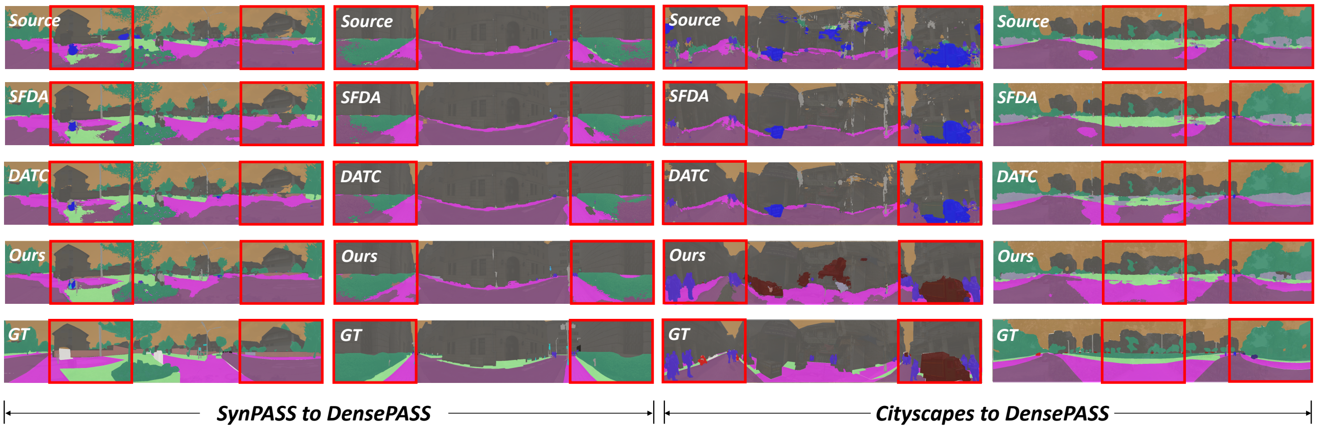

Figure 4. Example visualization results. (a) source, (b) SFDA [25], (c) DATC [41], (d) Ours, (e) Ground Truth (GT).

Method	SF	mIoU	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	Δ
Trans4PASS-T [49]	✗	53.18	48.54	16.91	79.58	65.33	55.76	84.63	59.05	37.61	-
Trans4PASS-S [49]	✗	55.22	48.85	23.36	81.02	67.31	69.53	86.13	60.85	39.09	-
DAFormer [17]	✗	54.67	49.69	25.15	77.70	63.06	65.61	86.68	65.12	48.13	-
DPPASS [59]	✗	55.30	52.09	29.40	79.19	58.73	47.24	86.48	66.60	38.11	-
DATR [58]	✗	56.81	54.62	29.50	80.03	67.35	63.75	87.67	67.57	37.10	-
Source w/ seg-b1	✓	38.65	40.93	10.89	67.67	36.86	15.56	26.43	42.68	27.16	-
SFDA w/ seg-b1 [25]	✓	42.70	41.65	8.46	69.97	47.48	33.24	72.01	47.61	32.77	+4.05
DTAC w/ seg-b1 [41]	✓	43.06	43.51	8.35	70.10	35.79	40.73	70.52	49.49	32.94	+4.41
Ours w/ seg-b1	✓	48.78	45.36	15.83	75.70	**49.16**	55.68	**82.07**	54.82	33.76	**+10.13**
Ours w/ seg-b2	✓	**50.12**	**49.92**	**27.22**	**76.22**	47.81	**64.13**	79.47	**56.83**	**35.76**	**+11.47**

Table 2. Experimental results of 8 selected categories in panoramic semantic segmentation on C-to-D. SF: Source-free UDA. The **bold** and underline denote the best and the second-best performance in source-free UDA methods, respectively.

3.4. Optimization

The training objective for learning the target model containing three losses is defined as:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{ppa} + \gamma \cdot \mathcal{L}_{cda} + \mathcal{L}_{bns} + \mathcal{L}_{sup} \qquad (12)$$

where $\mathcal{L}_{ppa}$ is the MSE loss from PPAM, $\mathcal{L}_{cda}$ refers to the KL loss from CDAM, $\mathcal{L}_{sup}$ denotes the CE loss for the prediction pseudo label supervision loss, $\mathcal{L}_{bns}$ refers to the BNS guided feature loss, and λ and γ are the trade-off weights of the proposed loss terms.

4. Experiments and Analysis

As the first SFUDA method for panoramic image segmentation, there is no prior method for direct comparison. We thus empirically validate our method by comparing it with the existing UDA and panoramic segmentation methods on three widely used benchmarks.

4.1. Datasets and Implementation Details.

Cityscapes [11] is a real-world dataset collected for autonomous driving that contains street scenes. DensePASS [27] is a panoramic dataset designed for capturing diverse street scenes. SynPASS [50] is a synthetic dataset consisting of 9080 synthetic panoramic images. Stanford2D3D [3] is an indoor panoramic dataset which has 1413 panoramic images. Overall, the experiments are conducted on both real-world (Cityscapes-to-DensePASS, C-to-D, and Stanford2D3D-pinhole-to-Stanford2D3D-panoramic, SPin-to-SPan) and synthetic-to-real (SynPASS-to-DensePASS, S-to-D) scenarios.

4.2. Experimental Results.

We first evaluate our proposed framework under the S-to-D scenario. The experimental results are shown in Tab. 1. Our proposed method consistently outperforms source-free UDA methods [25] and [41] and even achieves panoramic semantic segmentation performance closer to that of the UDA method Trans4PASS [50] which utilizes the source data

Method	SF	mIoU	Ceiling	Chair	Door	Floor	Sofa	Table	Wall	Window	Δ
PVT-S w/ MPA [49]	✗	57.95	85.85	51.76	18.39	90.78	35.93	65.43	75.00	40.43	-
Trans4PASS w/ MPA [49]	✗	64.52	85.08	58.72	34.97	91.12	46.25	71.72	77.58	50.75	-
Trans4PASS+ [50]	✗	63.73	90.63	62.30	24.79	92.62	35.73	73.16	78.74	51.78	-
Trans4PASS+ w/ MPA [50]	✗	67.16	90.04	64.04	42.89	91.74	38.34	71.45	81.24	57.54	-
SFDA [25]	✓	54.76	79.44	33.20	52.09	67.36	22.54	53.64	69.38	60.46	-
Ours w/ b1	✓	57.63	73.81	29.98	63.65	73.49	31.76	49.25	72.89	66.22	**+2.87**
Ours w/ b2	✓	65.75	82.88	38.00	**65.81**	86.71	36.32	66.10	80.29	**69.88**	**+10.99**

Table 3. Experimental results on indoor Stanford2D3D [3]. The **bold** denotes the best performance among UDA and SFUDA methods.

Loss Function Combinations					C-to-D		S-to-D	
$\mathcal{L}_{sup}$	$\mathcal{L}_{ppa}$	$\mathcal{L}_{sft}$	$\mathcal{L}_{cda}$	$\mathcal{L}_{bns}$	mIoU	Δ	mIoU	Δ
✓					38.65	-	35.81	-
✓	✓				45.42	+6.77	38.37	+2.56
✓	✓	✓			46.23	+7.58	38.49	+2.68
✓			✓		44.24	+5.59	38.38	+2.57
✓			✓	✓	44.79	+6.14	38.52	+2.71
✓	✓	✓	✓	✓	48.78	+10.13	41.78	+5.97

Table 4. Ablation study of different module combinations.

Combinations	τ_g+τ_p	τ_g+τ_f	τ_g+τ_p+τ_f
mIoU	44.14	44.28	**45.42**

Table 5. Ablation study of different prototype combinations.

in the adaptation procedure. Our proposed method brings significant performance gain of +3.57% and +3.54% with SegFormer-B1 backbone then SFDA [25] and DATC [41], respectively. We also provide the TSNE visualization in Fig. 5 (b) and qualitative results in Fig. 4. Apparently, our method gains a significant improvement in distinguishing the pixels in panoramic images in both prediction and high-level feature space. As shown in Tab. 2, we then evaluate our proposed framework under the C-to-D scenario. Our proposed method significantly outperforms source-free methods [25, 41] and some panoramic semantic segmentation methods [43, 46, 48]. Specifically, our method achieves a significant performance gain over SFDA [25] and DTAC [41] by +6.08% and +5.72%, respectively. This demonstrates that our proposed method endowed by PPAM and CDAM is more suitable for panoramic semantic segmentation tasks. Furthermore, as shown in the qualitative results in Fig. 4, our method achieves better segmentation in driving-related categories, such as rider and car.

We also provide TSNE visualizations [35] in Fig. 5 (a), showing that our proposed method brings significant improvements in distinguishing pixels from different categories in high-level feature space. Additionally, we evaluated our proposed method on the Stanford2D3D [3] dataset and compared it with the SFDA [25] and MPA [50] methods. As shown in the following table, our proposed method significantly outperforms the SFDA by +7.09% mIoU and is on par with the MPA method using source data (61.85% vs. 67.16%). Notably, for some categories, such as door (57.90% vs. 42.89%) and window (68.06% vs. 57.54%), our method event outperforms the MPA [50].

5. Ablation Study

Different Loss Function Combinations. To assess the effectiveness of the proposed modules, we conduct ablation experiments on both real-world and synthetic-to-real scenarios with various loss combinations. All of the proposed modules and loss functions have a positive impact on improving segmentation performance. Notably, our PPAM yields a significant performance gain of +6.77%. This indicates that PPAM alleviates the intricate semantics and distortion problem with the tangent, and our proposed FFP projection is valid. This is further supported by the qualitative results presented in Fig. 4. Additionally, our proposed CDAM achieves a performance gain of +5.59% compared to the source baseline, which means that CDAM imitates the spatial and channel-wise distributions of ERP and FFP features and further addresses the style discrepancy problems.

Ablation of Different Prototype Combinations. To validate the effectiveness of all the prototypes in PPAM, we conduct experiments on C-to-D using SegFormer-B1 and only $\mathcal{L}_{sup}$ and $\mathcal{L}_{ppa}$. The results of the performance with different prototype combinations are presented in Tab. 5. Both prototypes from TP and FFP have a positive effect on PPAM, with τ_p and τ_f resulting in mIoU improvements of +5.49% and +5.63%, respectively, compared to the source baseline. When both prototypes are combined together, there is a mIoU gain of +6.77%, indicating that their combination is better for prototype-level adaptation.

Dual Attention vs. Cross Dual Attention. The dual attention (DA) approach proposed in SFDA [25] aligns the spatial and channel characteristics of features between the fake source and target data. In contrast, our cross dual attention (CDA) approach aligns the distribution between different

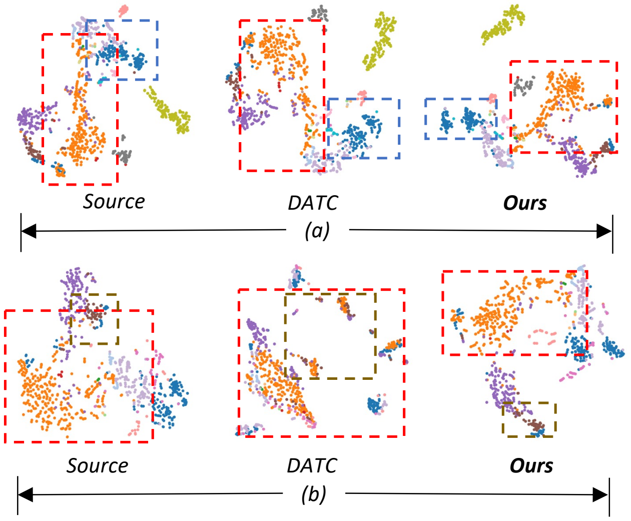

Figure 5. TSNE visualization of (a) Cityscapes-to-DensePASS and (b) SynPASS-to-DensePASS.

FoV	w/o	60°	72°	90°	120°	180°	360°
mIoU Δ	38.65 -	44.03 +5.38	44.16 +5.51	**44.28** **+5.63**	44.02 +5.37	41.65 +3.00	40.31 +1.66

Table 6. Ablation study of the FoV of our proposed FFP.

projections of the same spherical data, specifically ERP and FFP, resulting in more robust and stable knowledge transfer. Moreover, in our SFDA, we obtain spatial and channel characteristics across features, whereas DA operates within features. We also evaluate DA on the C-to-D scenario, and our CDA achieves 44.24% mIoU, while DA only reaches 41.53% mIoU. This indicates the proposed CDA is better for SFUDA in panoramic semantic segmentation.

Field-of-view of FFP. Most existing approaches for panoramic semantic segmentation, such as those proposed in [49, 50, 59], primarily focus on alleviating distortion by introducing distortion-aware components and distinct projection strategies. However, as discussed in Sec. 3.2, 360° images contain more intricate semantic information and object correspondence than the pinhole images, resulting in an obvious semantic mismatch between domains. Therefore, we propose the Fixed FoV Pooling (FFP) strategy to address the semantic mismatch. Experimental results show that the fixed FoV is the most influential factor in FFP, with an FoV of 90° achieving the best segmentation performance, as shown in Tab. 6, with a mIoU of 44.28%.

Ablation of Hyper-parameters. We now show the influence of hyperparameters γ and λ, which are the weights for the KL loss in CDAM and the MSE loss in PPAM, respectively. The experimental results are provided in Tab. 7.

Fine-tuning the Source Model. As the pre-trained model in the source (pinhole) domain is not an ideal model for the target (panoramic) image domain, we propose to fine-tune the source model with the loss function $\mathcal{L}_{sft}$, as described in Sec. 3.2. Tab. 4 demonstrates the effectiveness of the proposed $\mathcal{L}_{sft}$. When combined with the prototypical adaptation loss $\mathcal{L}_{ppa}$, adding $\mathcal{L}_{sft}$ results in a 6.77% mIoU gain compared with the source baseline of 38.65%. We present the performance metrics derived solely from the loss $\mathcal{L}_{sft}$ of PPAM: C-2-D registers at 44.94% while S-2-D records 36.74%. These results underscore the efficacy of $\mathcal{L}_{sft}$ integrated within our PPAM module. Concerning transfer-ability, our $\mathcal{L}_{sft}$ exhibits compatibility with various projection methods, *e.g.*, cube map. At its core, our fine-tuning loss seeks to align all projection images originating from the same panoramic source, irrespective of the employed projection technique. This intrinsic adaptability facilitates the application of $\mathcal{L}_{sft}$ across diverse projections. *More results refer to the supplementary material.*

γ	0	0.01	0.02	0.05	0.1	0.2
mIoU Δ	38.65 -	42.05 +3.40	43.24 +4.59	43.28 +4.63	**44.24** **+5.59**	43.07 +4.42

λ	0	50	60	80	100	120	150	200
mIoU Δ	38.65 -	43.13 +4.48	43.22 +4.57	45.36 +6.71	**45.42** **+6.77**	45.34 +6.69	45.33 +6.68	45.12 +6.47

Table 7. Ablation study of γ and λ.

6. Conclusion

In this paper, we investigated a new problem of achieving SFUDA for panoramic semantic segmentation. To this end, we proposed an end-to-end SFUDA framework to address the domain shifts, including semantic mismatch, distortion, and style discrepancies, between pinhole and panoramic domains.Experiments on both real-world and synthetic benchmarks show that our proposed framework outperforms prior approaches and is on par with the methods using source data.

Limitation and future work. One limitation of our proposed framework is the computational cost brought by the tangent projection during training, and there is still room for improvements in segmentation performance. However, components in our approach such as panoramic prototypes and fixed FoV projection have significant implications for the 360° vision, especially for the panoramic semantic segmentation. In the future, we plan to utilize the large language models (LLMs) and Multi-modal large language models (MLLMs) to alleviate the domain gaps, such as the semantic mismatches between pinhole and panoramic images.

Acknowledgement This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Ling Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022. 1

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

[3] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 6, 7

[4] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82:102617, 2022. 2

[5] Chaoqi Chen, Weiping Xie, Tingyang Xu, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, 2019. 2

[6] Jialei Chen, Daisuke Deguchi, Chenkai Zhang, Xu Zheng, and Hiroshi Murase. Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation. *Available at SSRN 4617170*. 2

[7] Jialei Chen, Chong Fu, Haoyu Xie, Xu Zheng, Rong Geng, and Chiu-Wing Sham. Uncertainty teacher with dense focal loss for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 149:106034, 2022.

[8] Jialei Chen, Daisuke Deguchi, Chenkai Zhang, Xu Zheng, and Hiroshi Murase. Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation. *arXiv preprint arXiv:2310.02296*, 2023.

[9] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2090–2099, 2019. 2

[10] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6829–6839, 2019. 2

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[12] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2020. 3

[13] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021. 2

[14] Xiaoqing Guo, Jie Liu, Tongliang Liu, and Yixuan Yuan. Simt: Handling open-set noise for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7032–7041, 2022. 2, 5

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

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[17] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 2, 6

[18] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021. 2, 5

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[21] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021. 2, 5

[22] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6929–6938, 2019. 2

[23] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2801–2810, 2022. 4

[24] Mengyi Liu, Shuhui Wang, Yulan Guo, Yuan He, and Hui Xue. Pano-sfmlearner: Self-supervised multi-task learning of depth and semantics in panoramic videos. *IEEE Signal Processing Letters*, 28:832–836, 2021. 2

[25] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2, 3, 5, 6, 7

[26] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2511, 2019. 2

[27] Chaoxiang Ma, Jiaming Zhang, Kailun Yang, Alina Roitberg, and Rainer Stiefelhagen. Densepass: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2766–2772. IEEE, 2021. 6

[28] Luke Melas-Kyriazi and Arjun K. Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12430–12440, 2021. 2

[29] Zak Murez, Soheil Kolouri, David J. Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2

[30] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. 2

[31] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser-Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 2

[32] Weifa Shen, Qixiong Wang, Hongxiang Jiang, Sen Li, and Jihao Yin. Unsupervised domain adaptation for semantic segmentation via self-supervision. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2747–2750. IEEE, 2021. 2

[33] Serban Stan and Mohammad Rostami. Unsupervised model adaptation for continual semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2593–2601, 2021. 2

[34] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 2

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[36] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 2

[37] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019. 2

[38] Qin Wang, Dengxin Dai, Lukas Hoyer, Olga Fink, and Luc Van Gool. Domain adaptive semantic segmentation with self-supervised depth estimation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8495–8505, 2021. 2

[39] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 5

[40] Haoyu Xie, Chong Fu, Xu Zheng, Yu Zheng, Chiu-Wing Sham, and Xingwei Wang. Adversarial co-training for semantic segmentation over medical images. *Computers in biology and medicine*, 157:106736, 2023. 2

[41] Cheng-Yu Yang, Yuan-Jhe Kuo, and Chiou-Ting Hsu. Source free domain adaptation for semantic segmentation via distribution transfer and adaptive class-balanced self-training. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2, 3, 5, 6, 7

[42] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, and Kaiwei Wang. Pass: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4171–4185, 2019. 1

[43] Kailun Yang, Xinxin Hu, Yicheng Fang, Kaiwei Wang, and Rainer Stiefelhagen. Omnisupervised omnidirectional semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1, 7

[44] Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. Source data-free unsupervised domain adaptation for semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2233–2242, 2021. 2, 3

[45] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021. 2

[46] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 1, 7

[47] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12612–12621, 2021. 2

[48] Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1, 7

[49] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16917–16927, 2022. 1, 2, 3, 4, 6, 7, 8

[50] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer

Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022. 1, 2, 4, 6, 7, 8

[51] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 2, 5

[52] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in neural information processing systems*, 32, 2019. 2

[53] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2039–2049, 2017. 2

[54] Yuyang Zhao, Zhun Zhong, Zhiming Luo, Gim Hee Lee, and Nicu Sebe. Source-free open compound domain adaptation in semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7019–7032, 2022. 2

[55] Xu Zheng, Chong Fu, Haoyu Xie, Jialei Chen, Xingwei Wang, and Chiu-Wing Sham. Uncertainty-aware deep co-training for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 149:106051, 2022. 2

[56] Xu Zheng, Yunhao Luo, Hao Wang, Chong Fu, and Lin Wang. Transformer-cnn cohort: Semi-supervised semantic segmentation by the best of both students. *arXiv preprint arXiv:2209.02178*, 2022.

[57] Xu Zheng, Yunhao Luo, Pengyuan Zhou, and Lin Wang. Distilling efficient vision transformers from cnns for semantic segmentation. *arXiv preprint arXiv:2310.07265*, 2023. 2

[58] Xu Zheng, Tianbo Pan, Yunhao Luo, and Lin Wang. Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18687–18698, 2023. 2, 6

[59] Xu Zheng, Jinjing Zhu, Yexin Liu, Zidong Cao, Chong Fu, and Lin Wang. Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. *arXiv preprint arXiv:2303.14360*, 2023. 1, 2, 4, 6, 8

[60] Jinjing Zhu, Yunhao Luo, Xu Zheng, Hao Wang, and Lin Wang. A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11720–11730, 2023. 2

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