360SFUDA++: Towards Source-free UDA for Panoramic Segmentation by Learning Reliable Category Prototypes

Xu Zheng, *Student Member, IEEE,* Pengyuan Zhou, Athanasios V. Vasilakos *Senior Member, IEEE,* Lin Wang*, *Member, IEEE,*

***Abstract*—In this paper, we address the challenging source-free unsupervised domain adaptation (SFUDA) for pinhole-to-panoramic semantic segmentation, given only a pinhole image pre-trained model (*i.e.*, source) and unlabeled panoramic images (*i.e.*, target). Tackling this problem is non-trivial due to three critical challenges: 1) semantic mismatches from the distinct Field-of-View (FoV) between domains, 2) style discrepancies inherent in the UDA problem, and 3) inevitable distortion of the panoramic images. To tackle these problems, we propose 360SFUDA++ that effectively extracts knowledge from the source pinhole model with only unlabeled panoramic images and transfers the reliable knowledge to the target panoramic domain. Specifically, we first utilize Tangent Projection (TP) as it has less distortion and meanwhile slits the equirectangular projection (ERP) to patches with fixed FoV projection (FFP) to mimic the pinhole images. Both projections are shown effective in extracting knowledge from the source model. However, as the distinct projections make it less possible to directly transfer knowledge between domains, we then propose Reliable Panoramic Prototype Adaptation Module (RP^{2}AM) to transfer knowledge at both prediction and prototype levels. RP^{2}AM selects the confident knowledge and integrates panoramic prototypes for reliable knowledge adaptation. Moreover, we introduce Cross-projection Dual Attention Module (CDAM), which better aligns the spatial and channel characteristics across projections at the feature level between domains. 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 360SFUDA++ achieves significantly better performance than prior SFUDA methods. Project Page: https://vlislab22.github.io/360SFUDA/**

***Index Terms*—Source-Free Unsupervised Domain Adaptation, Panoramic Semantic Segmentation**

I. Introduction

THE comprehensive scene perception capabilities afforded by 360° cameras have rendered them exceedingly popular for a diverse array of applications, including autonomous driving and virtual reality [1]. Unlike pinhole cameras, which capture 2D planar images within a confined field-of-view (FoV), 360° cameras provide an expansive FoV spanning 360°×180°. Consequently, there has been a concerted research endeavors towards panoramic semantic segmentation [2]–[6], aiming at achieving comprehensive scene understanding for intelligent systems.

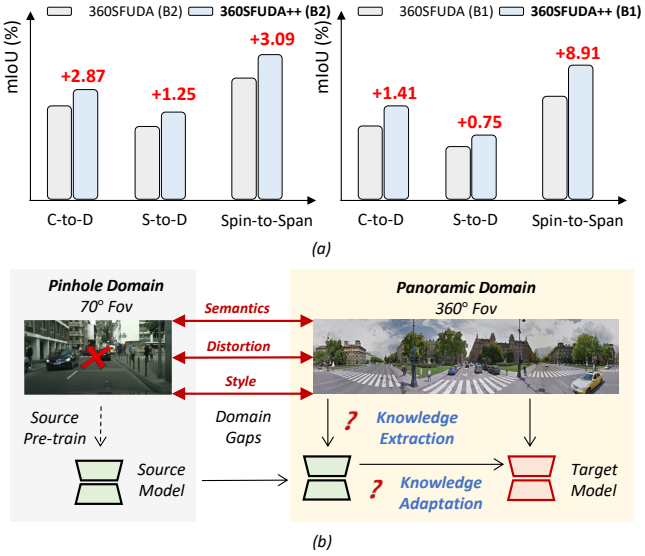

Fig. 1. (a) Performance comparison between [7] and 360SFUDA++; (b) Prototype comparison between [7] and 360SFUDA++ on outdoor C-to-D scenario, different colors stand for different categories.

Generally, data captured by 360° cameras in spherical form is commonly projected onto 2D planar representations, such as the Equirectangular Projection (ERP), to integrate seamlessly into existing imaging pipelines [1], while still preserving omni-directional information[1]. However, ERP images always suffer from the inevitable distortion and object deformation due to the non-uniformly distributed pixels [8]. Meanwhile, learning effective panoramic segmentation models is often hindered by the lack of large precisely labeled datasets due to the difficulty of annotation. For these reasons, some unsupervised domain adaptation (UDA) methods [4], [8], [9] 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. A notable example is the recent large-scale model, SAM [10], which has made

* Corresponding Author
Xu Zheng is with the AI Thrust, HKUST(GZ), Guangdong, China. E-mail: zhengxu128@gmail.com. Pengyuan Zhou is with the Aarhus University, Denmark. E-mail:pengyuan.zhou@ece.au.dk. Athanasios V. Vasilakos is with the University of Agder, Norway. E-mail: th.vasilakos@gmail.com. Lin Wang is with AI/CMA Thrust, HKUST(GZ) and Dept. of CSE, HKUST, Hong Kong SAR, China, E-mail: linwang@ust.hk.

[1]Throughout this paper, the terms "omnidirectional" and "panoramic" are used interchangeably, with ERP images often denoting panoramic images.

significant advancements in instance segmentation for pinhole images. However, the size of the source datasets (10TB) renders them impractical for reuse in downstream tasks [10].

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 (b), different from existing SFUDA methods, *e.g.*, [11]–[13] for the pinhole-to-pinhole image adaptation, transferring knowledge from the pinhole-to-panoramic image domain is hampered by: **1)** semantic mismatches arising from the disparate FoV between pinhole and 360° cameras, namely 70° vs. 360°; **2)** inevitable inherent distortion of the ERP; **3)** style discrepancies caused by the distinct camera sensors and captured scenes. In Tab. I, we demonstrate that simply applying existing SFUDA methods to the specific pinhole-to-panoramic adaptation problem yields only marginal performance enhancements.

To this end, we propose 360SFUDA++ that effectively extracts knowledge from the source pinhole model with only unlabeled panoramic images and transfers the reliable knowledge to the target panoramic domain. *Our key idea is to leverage the multi-projection versatility of 360° data for efficient and reliable domain knowledge transfer*. **360SFUDA++** enjoys two key technical contributions. Specifically, we leverage Tangent Projection (TP)[2] and divide the ERP images into patches with a fixed FoV, dubbed Fixed FoV Projection (FFP), to mimic pinhole images and extract knowledge from the source model, aiming at alleviating distortion and semantic mismatches between domains. Both projection techniques enable efficient knowledge extraction from the source model. Considering the distinct projections make it hardly possible to directly transfer the extracted knowledge to the target model, we introduce the reliable panoramic prototype adaptation module (RP^{2}AM) (Sec. III-B1) to achieve knowledge transfer at the prototype level (See Fig. 1). RP^{2}AM employs cross-projection prediction assessment (Sec. III-B2) to distinguish confident and uncertain pixels across three projections, extracting prototypes in the target panoramic domain (Sec. III-B3). The global prototypes from the source model with TP and FFP images are updated iteratively during adaptation to enhance knowledge extraction. Additionally, RP^{2}AM fine-tunes the source model using prototypes from FFP images to improve knowledge extraction (Sec. III-B4). Aligning prototypes from each FFP image enhances the model's awareness of distortion and semantics across the FoV.

For efficient knowledge adaptation, we apply prediction-level and prototype-level loss constraints to facilitate knowledge transfer to the unlabeled target panoramic domain (Sec. III-C). FFP predictions are reconstructed as pseudo labels for the target model. Prototype-level loss constraints are enforced in both intra-projection and cross-projection prototypical adaptation. Intra-projection prototypical adaptation occurs between confident and uncertain prototypes obtained with ERP images. Cross-projection prototypical adaptation utilizes global panoramic prototypes extracted from TP and FFP images from the source model, continually updated throughout the adaptation process to supervise confident prototypes in the target domain.

Moreover, the knowledge derived from the source model goes beyond predictions and prototypes, encompassing high-level features that encapsulate crucial image characteristics, significantly enhancing the performance of the target model. To address this, we introduce the Cross-projection Dual Attention Module (CDAM) (Sec. III-C), aiming at harmonizing spatial and channel characteristics across projections and domains, maximizing the utilization of knowledge from the source model and mitigating the style discrepancy problems. CDAM reconstructs source model features from FFP images to provide a panoramic understanding of the surrounding environment, aligning them with ERP features from the target model to facilitate effective knowledge transfer.

We conduct extensive experiments on both synthetic and real-world benchmarks, including outdoor and indoor scenarios. We adapt the state-of-the-art (SoTA) SFUDA methods [11], [13]–[17] – designed for pinhole-to-pinhole image adaptation – to our problem in addressing the panoramic semantic segmentation. As shown in Fig. 1, the results show that our framework significantly outperforms these methods by large margins of +1.25%, +2.87%, and +3.09% on three benchmarks. We also evaluate our method against UDA methods [4], [8], [9], [18], using source data, the results demonstrate its comparable performance.

In summary, our main contributions are as follows: **(I)** we propose 360SFUDA++, which incorporates RP^{2}AM and CDAM to address the challenging source-free UDA task from pinhole to panoramic images; **(II)** The RP^{2}AM is proposed to achieve the panoramic knowledge adaptation with reliable prototypes between domains; **(III)** CDAM aims to harmonize spatial and channel characteristics across different projections of sphere data, aiming at mitigating the style discrepancy problem. **(IV)** Extensive experiments are conducted on both synthetic and real-world benchmarks (including indoor and outdoor scenarios) demonstrate the superiority of our 360SFUDA++.

This work is an improvement over our CVPR 2024 work [7], achieved by substantially extending the method and experiment in the following ways. **(I)** We introduced the cross-projection pixel-wise prediction assessment in RP^{2}AM to discern confident and uncertain predictions for better knowledge adaptation (Sec. III-B2); **(II)** We upgraded the prototype extraction to cross-projection prototype extraction to obtain confident and uncertain prototypes for better knowledge transfer (Sec. III-B3); **(III)** We added the loss constraints between uncertain and confident prototypes to achieve intra-projection knowledge adaptation (Sec. III-B3); **(IV)** We aligned the confident prototypes with the global panoramic prototypes for accurate and reliable prototypical knowledge adaptation (Sec. III-C); **(V)** We conducted more comparison experiments between the new framework and the previous version, as well as the existing SoTA SFUDA methods (Sec. IV-B); **(VI)** We validated the effectiveness of the introduced strategies and components with extensive quantitative and qualitative analysis (Sec. V).

[2]See Algorithm 2 in the supplementary material for tangent projection.

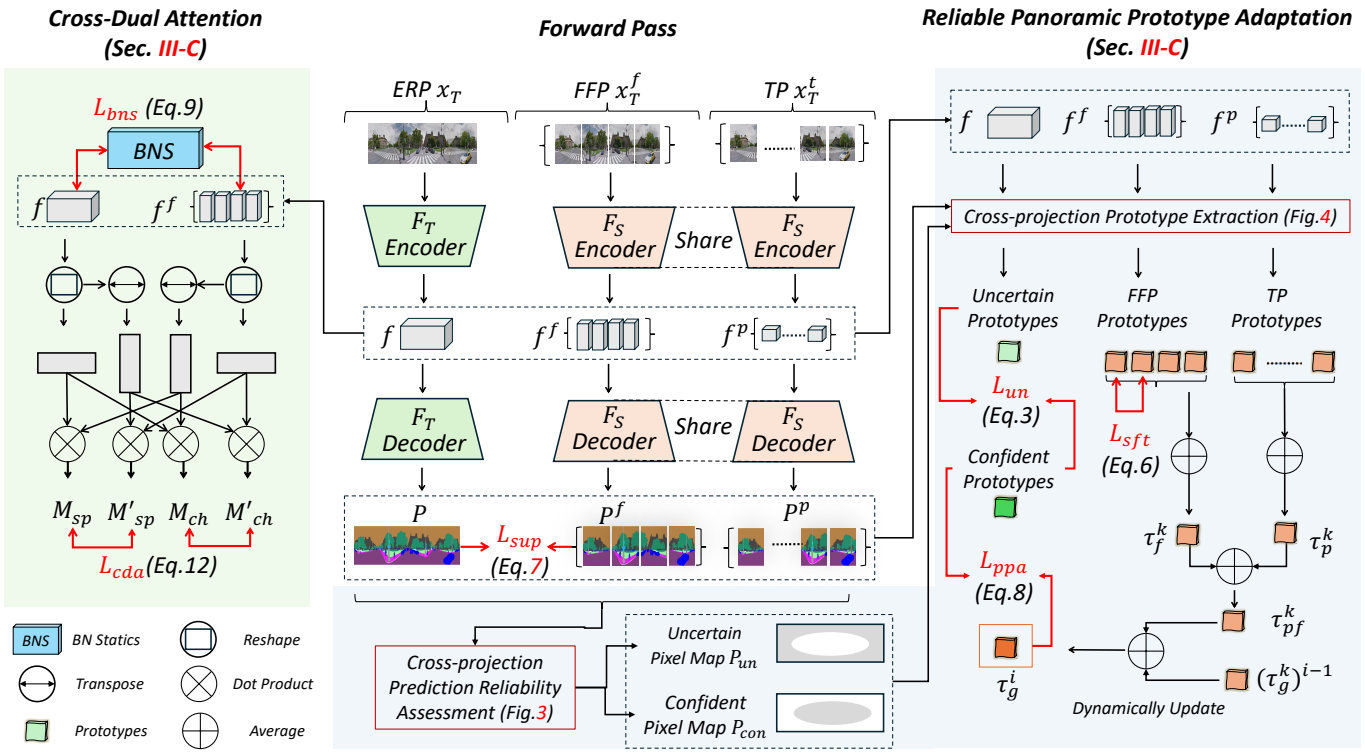

Fig. 2. Overall framework of our proposed 360SFUDA++.

II. Related Work

A. 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 [19]–[38]. However, the source domain data may not always be accessible due to the privacy protection and data storage concerns. Intuitively, source-free UDA (SFUDA) [15], [16], [39], [40] methods are proposed to adapt source models to a target domain without access to the source data. Existing source-free UDA methods can be categorized into two distinct approaches: those generating images independently of the source data, and those employing self-supervision techniques with target pseudo-labels. [11]–[13], [15], [41]–[44]. For instance, [43] introduces a dual-branch collaborative learning framework for SFUDA which separate target data into confident samples and uncertain samples to figure out unreliable pixel-wise segmentation. In this paper, we attempt to achieve SFUDA from the pinhole image domain to the panoramic image domain. This task is nontrivial to be tackled due to the semantic mismatches, style discrepancies, and inevitable distortion of panoramic images [7]. Unlike these methods that focus on the source domain data estimation [11], [12], we propose 360SFUDA++, a novel SFUDA framework that effectively extracts knowledge from the source model with only panoramic images and transfers the knowledge to the target panoramic image domain. Different from the prior prototypical adaptation methods, such as [43] which only focuses the confident and unreliable problem at sample level, our 360SFUDA++ subtly utilize the multi-projection versatility of panoramic data and achieve reliable prototypical adaptation in the pixel-level. Experiments also show that naively applying these methods leads to less optimal performance (See Tab. I).

B. UDA for Panoramic Semantic Segmentation

It can be classified into three types, including adversarial training [8], [45]–[48], pseudo labeling [49]–[52] and prototypical adaptation methods [4], [9]. Specifically, the first line of research applies alignment approaches to capture the domain invariant characteristics of images [45], [53], [54], feature [8], [45], [55], [56] and predictions [57], [58]. The second type of method generates pseudo labels for the target domain training. The last line of research, *e.g.*, Mutual Prototype Adaption (MPA) [4], utilizes the prototypical adaptation and mutually aligns the high-level features with the prototypes between domains. However, these methodologies often treat panoramic images as pinhole images during prototype extraction and adaptation, thereby neglecting the nuanced semantic details, object correspondences, and distortion inherent in panoramic FoV. In this paper, we address the SFUDA problem for panoramic segmentation by improving our previous work 360SFUDA [7]. Considering the distinct projection discrepancies between source and target domains, we propose an RP^{2}AM to transfer knowledge at the prototype level. Differently, we take the cross-projection pixel-wise predictions' confidence into consideration and further extract reliable prototypes for better knowledge adaptation.

III. Methodology

A. Overview

The overall framework of our 360SFUDA++ is shown in Fig. 2. Given only the source model F_S and unlabeled

panoramic image data D_T, the SFUDA objective is to learn a target model F_T that effectively transfers knowledge from F_S to the common K categories across both domains. Different from the traditional pinhole image-to-image adaptation [11]–[13], the knowledge adaptation from pinhole to panoramic image domains is notably challenged by three primary factors, specifically: 1) semantic mismatch between pinhole and panoramic images caused by the FoV variations (70° vs. 360°), which means there are more objects and cross-object correlations in panoramic images with 360° FoV; 2) inevitable distortion and deformation in ERP of panoramic images; and 3) ubiquitous style discrepancies in almost all UDA tasks.

Therefore, naively applying existing SFUDA methods exhibits sub-optimal segmentation performance (See Tab. I), while UDA methods with source data, *e.g.*, [11] for panoramic segmentation fail to address the semantic discrepancies between the pinhole and panoramic images. Intuitively, the key challenges are : **1)** how to extract knowledge from the source pinhole model with only unlabeled panoramic images and **2)** how to efficiently transfer knowledge to the target panoramic image domain. Overall, in our 360SFUDA++, **the key idea** *is to leverage the multi-projection versatility of 360° data for efficient domain knowledge transfer.*

To tackle the first challenge (Sec. III-B), we utilize the Tangent Projection (TP) which is distinguished by its ability to mitigate distortion problems more effectively than ERP images [59], leveraging the reduced distortion characteristics of TP to extract knowledge from the source pinhole model. Concurrently, we segment ERP images into discrete patches, each possessing a constant FoV to mimic the pinhole images, dubbed Fixed FoV Projection (FFP). Both TP and FFP facilitate the effective extraction of knowledge from the source model. However, the inherent differences in projection formats preclude a direct transfer of knowledge between domains. To overcome this obstacle, we introduce the Reliable Panoramic Prototype Adaptation Module (RP^{2}AM), designed to generate panoramic prototypes suitable for knowledge adaptation. To address the second challenge (Sec. III-C), 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 between pinhole and panoramic images.

B. Knowledge Extraction

As illustrated in Fig. 2, given the unlabled 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 $D_T^f = \{x_T^f | x_T^f \in \mathbf{R}^{H\times W/4\times 3}\}$. Note that one ERP image corresponds to 18 TP images as [8], [60] and 4 FFP images with a fixed FoV of 90° (ablation can be found in Sec. V-B). To obtain the high-level features and predictions from the source model for knowledge adaptation, TP and FFP 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 domain 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 $P, f = F_T(x_T)$. To overcome inherent differences caused by the projection formats, we introduce the Reliable Panoramic Prototype Adaptation Module (RP^{2}AM) to obtain panoramic prototypes for effective and reliable knowledge adaptation.

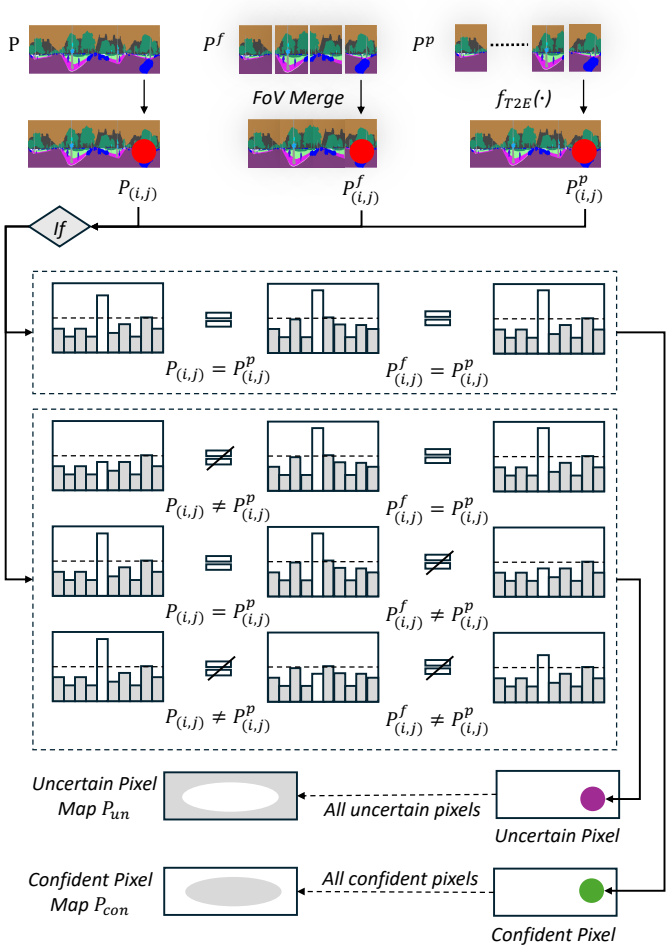

Fig. 3. Illustration of the cross-projection pixel-wise prediction assessment.

1) Reliable Panoramic Prototype Adaptation Module: In contrast to the previous version of this work [7], namely panoramic prototype adaptation module (PPAM), which posses three distinct characteristics, including: (a) class-wise prototypes are obtained from distinct projections directly to achieve knowledge transfer; (b) global prototypes are iteratively updated across training; and (c) hard pseudo labels are softened in the feature space to take full use of the source knowledge. Deviating from the previous work [7], we put forward a better prototypical adaptation strategy with more reliable knowledge extraction and transfer.

Compared to prior UDA and SFUDA methods using prototypical adaptation, *e.g.*, CFA [18], MPA [4], [9] and our previous PPAM [7], our RP^{2}AM possesses two distinct characteristics: **(a)** cross-projection prototypical adaptation is conducted between ERP and the other two projections, *i.e.*, TP and FFP, aiming at alleviating the distortion and semantic mismatch problems; and **(b)** intra-projection prototypical adaptation is

performed with ERP and FFP, aiming to achieve reliable prototypical adaptation for the ERP and distortion-aware ability for the FFP. Specifically, we first project the source model predictions P^p, P^f into pseudo labels:

$$\hat{y}^p_{(h,w,k)} = 1_{k \doteq argmax(P^p_{h,w,:})},$$
$$\hat{y}^f_{(H,W/4,k)} = 1_{k \doteq argmax(P^f_{H,W/4,:})}, \quad (2)$$

where k denotes the semantic category. Subsequently, we propose the cross-projection pixel-wise prediction assessment to select the confident predicted pixels across three projections.

2) Cross-projection Pixel-wise Prediction Assessment: As shown in Fig. 3, we transfer all the prediction maps from different projections into the ERP image (*e.g.*, 400 × 2048 in DensePASS) by merging the FFP and applying transformation function $f_{T2E}(\cdot)$ as [8] to TP images. Then, for each pixel coordinate (i, j), a comparative evaluation of the prediction logits $P_{(i,j)}$, $P^f_{(i,j)}$, and $P^p_{(i,j)}$ is conducted to distinguish confident from uncertain pixel predictions. Specifically, a pixel (i, j) is deemed confident if the condition $P_{(i,j)} = P^f_{(i,j)} = P^p_{(i,j)}$ holds, signifying concordance in the prediction outputs across the varied projections for $P_{(i,j)}$. This allows for building up a confident pixel map from pixels, such as $P_{(i,j)}$. Conversely, the pixels with different predictions are defined as uncertain pixels, facilitating the generation of an uncertain pixel map.

3) Cross-projection Prototype Extraction: We then utilize the predicted maps as masks to get the confident and uncertain predictions P_{con} and P_{un} which are used to attain the confident and uncertain prototypes for ERP images. P_{con} and P_{un} are first converted to pseudo labels as the same operations in Eq. 2. Then we up-sample the ERP features f to fit the spatial size as P_{con} and P_{un}. The uncertain and confident prototypes τ_{un} and τ_{con} for ERP images are achieved by masked average pooling (MAP) operation, as shown in Fig. 4. To realize the intro-projection prototypical knowledge adaptation from the reliable pixels to the uncertain pixels, the Mean Squared Error (MSE) loss is imposed between the reliable and uncertain prototypes as follows:

$$\mathcal{L}_{un} = \sum_{\alpha \neq \beta}^{4} \{ \frac{1}{K} \sum_{k \in K} (\tau_{con} - \tau_{un})^2 \}. \quad (3)$$

Subsequently, class-specific masked features are derived by amalgamating the up-sampled features with their corresponding pseudo labels, denoted as $\hat{y}^p_{(h,w,k)}$ for TP and $\hat{y}^f_{(H,W/4,k)}$ for FFP images. It is noteworthy that the prototypes for TP and FFP images, represented as $\sum_{a=1}^{18}(\tau^k_p)_a$ and $\sum_{b=1}^{4}(\tau^k_f)_b$ respectively, are acquired through the MAP operation, as depicted in Fig. 4. Within each projection, RP^{2}AM first integrates the prototypes:

$$\tau^k_p = avg(\sum_{a=1}^{18}(\tau^k_p)_a), \qquad \tau^k_f = avg(\sum_{b=1}^{4}(\tau^k_f)_b). \quad (4)$$

As shown in Fig. 2, τ^k_p and τ^k_f are integrated together as τ^k_{pf} to retain the less distortion characteristics inherent τ^k_p and the similar scale semantics of τ^k_f. The τ^k_{pf} is then used to update the panoramic global prototype τ^k_g, which is iteratively

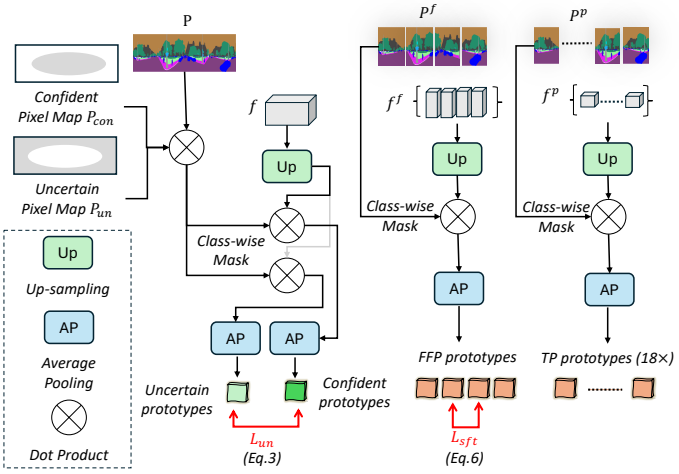

Fig. 4. Illustration of the cross-projection prototype extraction.

updated with τ^k_{pf}. To obtain more accurate and reliable prototypes, we update τ^k_g and τ^k_{pf} as follows:

$$\tau^i_g = \frac{1}{i}(\tau^k_{pf})^i + (1 - \frac{1}{i})(\tau^k_g)^{i-1}, \quad (5)$$

where $(\tau^k_g)^i$ and $(\tau^k_{pf})^i$ are the prototypes for category k in the $i-$th training epoch, $(\tau^k_g)^{i-1}$ is the panoramic global prototype saved in the last training epoch, i is the current epoch number. The panoramic global prototype τ^k_g is then used to give supervision for the confident prototype τ_{con} obtained from Sec. III-B2, as shown in Eq. 8.

4) Fine-tuning Source Model for Better Knowledge Extraction: Besides extracting prototypical knowledge from the source model, RP^{2}AM also fine-tunes the source model to improve the effectiveness of knowledge extraction. Given that each ERP image is converted to four FFP image patches with the same FoV, the source model consequently extracts four distinct sets of FFP features, denoted as f_f. Considering that these features originate from the same ERP image, we synchronize the class-wise prototypes across each set of FFP features, thereby enhancing the source model's overall performance and distortion awareness. Specifically, the prototypes $\sum_{\alpha=1}^{4}\tau_\alpha$ of the four FFP features are obtained through the same operations with τ^t_g.

Each FFP image encapsulates a unique, non-overlapping 90° FoV, presenting distinct levels of distortion and overlapping semantic content across the different FFP images. The alignment of prototypes corresponding to each FFP image not only bolsters the source model's capability to recognize and adapt to the distortion but also facilitates the exploration of complementary semantic content inherent within each FFP image patch. Thus, the MSE loss is imposed between each two of the prototypes, formulated as follows:

$$\mathcal{L}_{sft} = \sum_{\alpha \neq \beta}^{4} \{ \frac{1}{K} \sum_{k \in K} ((\tau^k_f)_\alpha - (\tau^k_f)_\beta)^2 \}. \quad (6)$$

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

C. Knowledge Adaptation

To facilitate the adaptation of knowledge to the target domain, we first enforce loss constraints on both predictions and prototypes. Moreover, we introduce a Cross-Dual Attention Module (**CDAM**) at the feature level. This module is designed to *enhance the alignment of both spatial and channel characteristics across domains and projections*. The CDAM leverages attention mechanisms [11] to adjust and synchronize feature representations, thereby improving the congruence between the source and target domains. This strategic incorporation of spatial and channel attention statistics within CDAM bridges the domain-specific discrepancies, ensuring a more effective and nuanced knowledge transfer.

Specifically, the FFP image patches are stitched to reconstruct an ERP image. It is then passed to the source model F_S to predict a pseudo label, which is employed to supervise the ERP predictions of the target model F_T. For simplicity, we use the Cross-Entropy (CE) loss, which is formulated as:

$$\mathcal{L}_{sup} = CE(P, 1_{k \doteq argmax(\{Rebuild(P^f_{H,W/4,:})\})}). \quad (7)$$

The prototype-level knowledge transfer loss is achieved by MSE loss between the panoramic global prototype τ^k_g and the confident prototype τ^k_{con} from ERP images:

$$\mathcal{L}_{ppa} = \frac{1}{K}\sum_{k\in K}(\tau^k_g - \tau^k_{con})^2. \quad (8)$$

With loss $\mathcal{L}_{ppa}$, the reliable prototypes are pushed together to transfer the knowledge extracted from the source domain to the target domain. In summary, with the proposed RP^{2}AM, we can effectively address the distortion and semantic mismatch problems at the prediction and prototype levels. We now tackle the style discrepancy problem at the feature level.

Cross-projection Dual Attention Module (CDAM). Inspired by the dual attention [11], focusing on spatial and channel characteristics, our CDAM imitates the spatial and channel-wise distributions of cross-projection features to alleviate the style discrepancies in UDA for panoramic semantic segmentation. Different from the basic dual attention in [11] which minimizes the distribution distance of the dual attention maps between the fake source and target data, our CDAM focuses on directly aligning the distribution between FFP and ERP of the same panoramic images rather than introducing additional frameworks or parameters to estimate fake source data as proxy. 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:

$$\begin{aligned}\mathcal{L}_{bns} =& ||\mu(F) - \bar{\mu}||^2_2 + ||\sigma^2(F) - \bar{\sigma}^2||^2_2 \\ &+ ||\mu(F') - \bar{\mu}||^2_2 + ||\sigma^2(F') - \bar{\sigma}^2||^2_2, \end{aligned} \quad (9)$$

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, 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:

$$\begin{aligned}\{M_{sp}\}_{ji} &= \frac{exp(f'_{[i:]} \cdot f^T_{[:j]})}{\sum^N_i exp(f'_{[i:]} \cdot f^T_{[:j]})}, \\ \{M'_{sp}\}_{ji} &= \frac{exp(f_{[i:]} \cdot f'^T_{[:j]})}{\sum^N_i exp(f_{[i:]} \cdot f'^T_{[:j]})}, \end{aligned} \quad (10)$$

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:

$$\begin{aligned}\{M_{ch}\}_{ji} &= \frac{exp(f'^T_{[i:]} \cdot f_{[:j]})}{\sum^C_i exp(f'_{[i:]} \cdot f^T_{[:j]})}, \\ \{M'_{ch}\}_{ji} &= \frac{exp(f^T_{[i:]} \cdot f'_{[:j]})}{\sum^C_i exp(f^T_{[i:]} \cdot f'_{[:j]})}. \end{aligned} \quad (11)$$

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

$$\mathcal{L}_{cda} = KL(M_{sp}, M'_{sp}) + KL(M_{ch}, M'_{ch}) \quad (12)$$

Algorithm 1 Framework of 360SFUDA++

Input: Pre-trained Source Model F_S; Unlabeled Panoramic Images (ERP) x_T, FFP image patches x^f_T, and TP image patches x^t_T;
Output: Target Model F_T;
Initialize F_T randomly;
for *each epoch* **do**
  for *each iteration* **do**
    Sample x_T and process x^f_T and x^t_T
    Forward Pass of F_S (Eq.1)
    Obtain pseudo labels (Eq.2)
    Cross-projection prediction assessment (Fig. 3)
    Prototype extraction (Fig. 4, Eq.4)
    $\mathcal{L}_{un}$ between uncertain and confident prototypes (Eq.3)
    Update global panoramic prototypes (Eq.5)
    Prototypical adaptation loss $\mathcal{L}_{ppa}$ (Eq.8)
    Fine-tuning F_S with $\mathcal{L}_{sft}$ (Eq.6)
    Align F and F' with $\mathcal{L}_{bns}$ (Eq.9)
    Feature level knowledge transfer (Eq.12)
  end
end
Final target model F_T.

D. Optimization

The overall optimization procedure is shown in Algorithm. 1. The training objective for learning the target model containing three losses is defined as:

$$\mathcal{L} = \mathcal{L}_{un} + \mathcal{L}_{ppa} + \gamma \cdot \mathcal{L}_{cda} + \mathcal{L}_{bns} + \mathcal{L}_{sup} \quad (13)$$

where $\mathcal{L}_{ppa}$ is the MSE loss from RP^{2}AM, $\mathcal{L}_{cda}$ refers to the KL loss from CDAM, $\mathcal{L}_{sup}$ denotes the CE loss for the

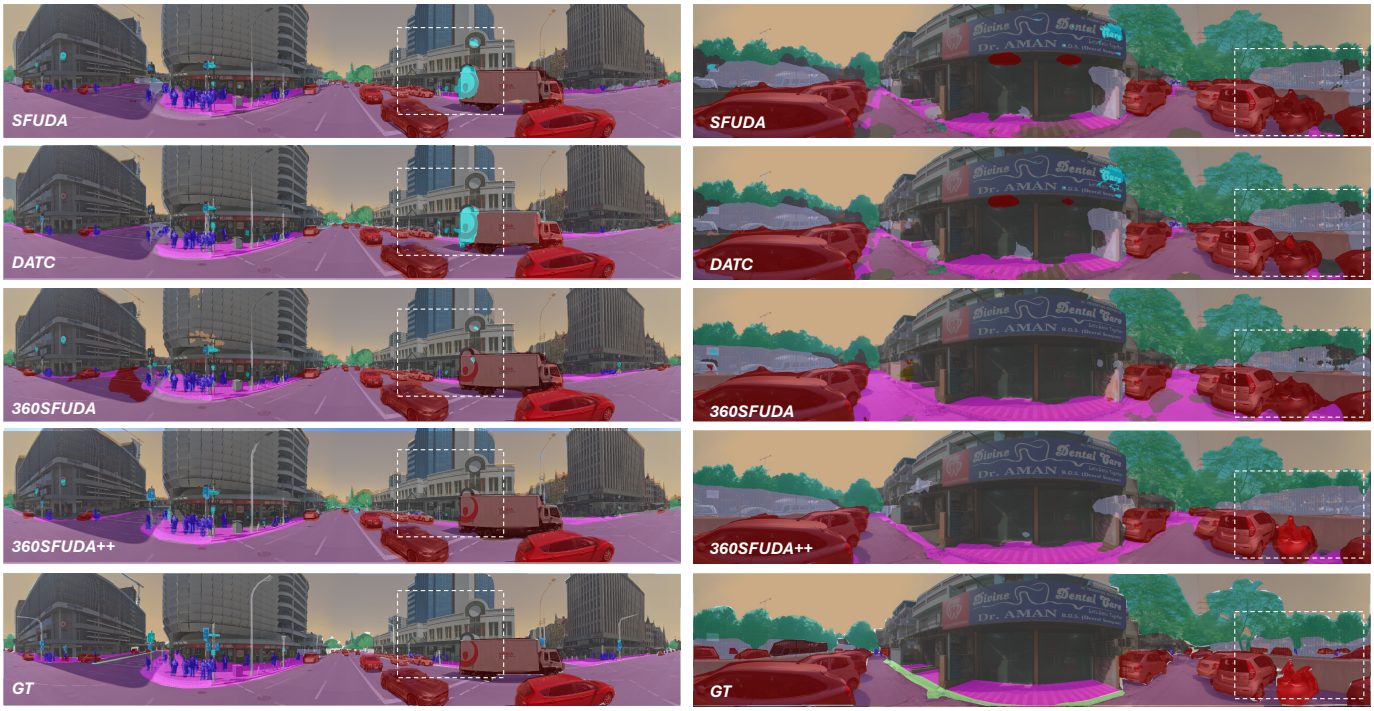

Fig. 5. Visualization results on C-to-D scenario. (a) source, (b) SFDA [11], (c) DATC [13], (d) 360SFUDA, (e) 360SFUDA++ (f) Ground Truth (GT).

TABLE I
EXPERIMENTAL RESULTS OF 19 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.

Method	SF	mIoU	Road	S.W.	Build.	Wall	Fence	Pole	Light	Sign	Vegt.	Terr.	Sky	Pers.	Rider	Car	Truck	Bus	Train	Motor	Bike
ECANet [61] (O.s.)	✗	43.02	81.60	19.46	81.00	32.02	39.47	25.54	3.85	17.38	79.01	39.75	94.60	46.39	12.98	81.96	49.25	28.29	0.00	55.36	29.47
P2PDA [5]	✗	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
SIM [62] (S.t.)	✗	44.58	68.16	32.59	80.58	25.68	31.38	23.60	19.39	14.09	72.65	26.41	87.88	41.74	16.09	73.56	47.08	42.81	56.35	47.72	39.30
PCS [6]	✗	53.83	78.10	46.24	86.24	30.33	45.78	34.04	22.74	13.00	79.98	33.07	93.44	47.69	22.53	79.20	61.59	67.09	83.26	58.68	39.80
DAFormer [22]	✗	54.67	73.75	27.34	86.35	35.88	45.56	36.28	25.53	10.65	79.87	41.64	94.74	49.69	25.15	77.70	63.06	65.61	86.68	65.12	48.13
Trans4PASS-T [4]	✗	53.18	78.13	41.19	85.93	29.88	37.02	32.54	21.59	18.94	78.67	45.20	93.88	48.54	16.91	79.58	65.33	55.76	84.63	59.05	37.61
DPPASS-T [8]	✗	55.30	78.74	46.29	87.47	48.62	40.47	35.38	24.97	17.39	79.23	40.85	93.49	52.09	29.40	79.19	58.73	47.24	86.48	66.60	38.11
DATR-S [18]	✗	56.81	80.63	51.77	87.80	44.94	43.73	37.23	25.66	21.00	78.61	26.68	93.77	54.62	29.50	80.03	67.35	63.75	87.67	67.57	37.10
Source	✓	38.65	65.26	29.40	77.04	15.14	28.72	14.15	9.36	10.55	69.09	21.10	82.91	40.98	10.42	68.56	32.90	44.94	50.98	37.74	25.19
SFDA [11]	✓	42.70	68.75	31.59	80.99	19.61	29.60	18.67	7.7	14.08	73.74	24.91	88.38	41.66	8.46	69.97	47.48	33.25	72.02	47.62	32.77
DATC [13]	✓	43.06	70.21	35.87	80.60	21.42	28.14	19.10	5.79	15.10	72.76	27.42	88.14	41.65	10.29	72.32	47.80	21.97	80.91	46.65	32.01
SFUDA [43]	✓	43.40	68.64	44.55	79.30	32.44	30.48	21.70	13.35	16.84	75.26	24.74	86.70	38.80	8.97	75.39	40.90	4.99	**85.24**	48.28	27.94
360SFUDA [7] w/ b1	✓	48.78	72.90	**48.10**	82.77	22.19	39.69	26.66	17.90	14.35	74.98	25.95	88.95	45.36	15.83	75.70	49.16	55.68	82.07	54.82	33.76
360SFUDA [7] w/ b2	✓	50.12	72.29	43.04	84.48	29.72	37.68	22.83	9.52	14.45	75.26	**34.53**	91.12	49.92	27.22	76.22	47.81	**64.13**	79.47	56.83	**35.76**
360SFUDA++ w/ b1	✓	50.19	68.59	46.70	84.06	29.08	34.74	**31.28**	20.31	16.84	75.04	23.07	92.20	50.03	18.32	78.75	**56.53**	49.90	83.63	59.00	35.46
360SFUDA++ w/ b2	✓	**52.99**	**74.00**	48.03	**85.86**	**34.41**	**41.28**	31.25	**22.59**	**17.41**	**76.74**	26.24	**92.65**	**54.60**	**30.76**	**79.41**	55.60	59.84	81.25	**59.89**	35.02

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.

IV. EXPERIMENTS AND ANALYSIS

We empirically validate our method by comparing it with the existing SFUDA, UDA, and panoramic segmentation methods on three widely used benchmarks.

A. Datasets and Implementation Details.

(1) Cityscapes-to-DensePASS **(C-to-D):** Cityscapes [65] constitutes a comprehensive real-world dataset meticulously collected for autonomous driving applications, encompassing street scenes sourced from 50 distinct cities. It offers precise pixel-wise annotations across 19 semantic categories. The dataset's official split includes 2975 images for training and 500 images for validation. In this study, we leverage the official training set (2975 images) as the source domain data for acquiring the pre-trained source model. DensePASS [66], on the other hand, is a panoramic dataset tailored to capture diverse street scenes worldwide, comprising 2500 panoramas. Within this collection, a subset of 100 panoramas has been meticulously annotated with a high degree of precision, targeting categories that are crucial for navigation. This subset is

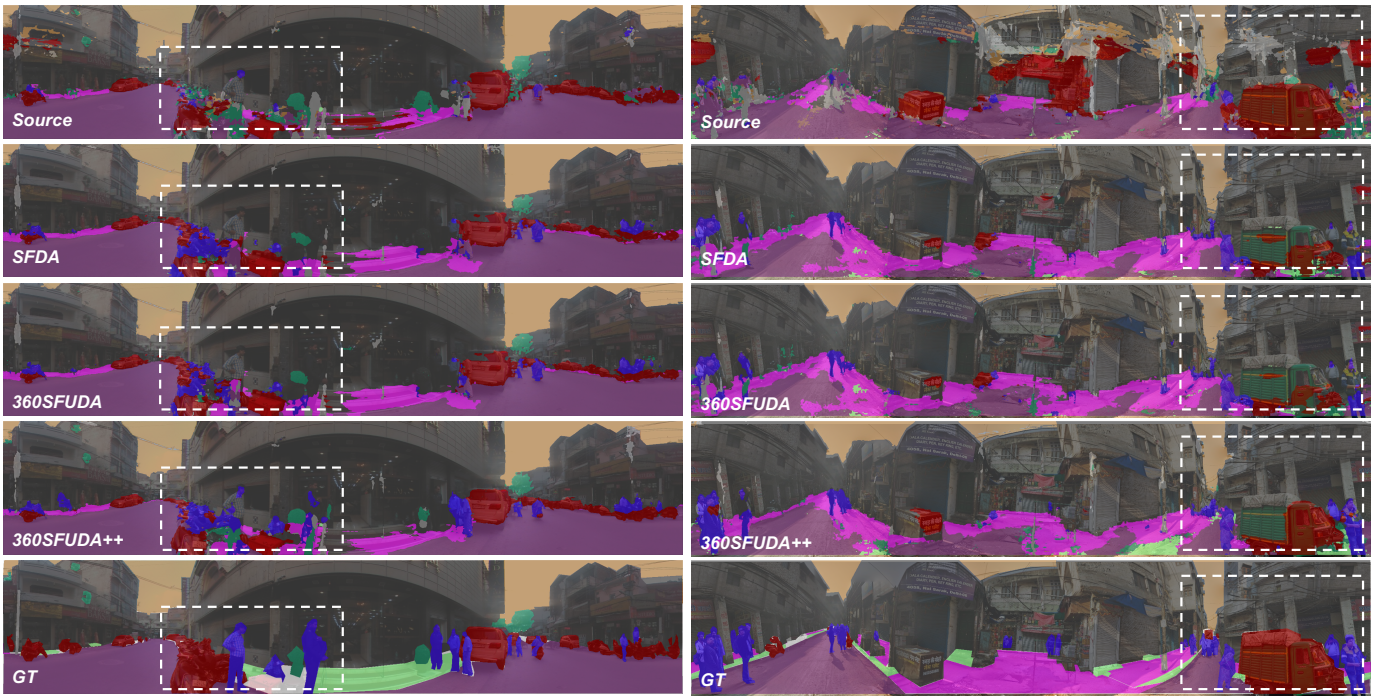

Fig. 6. Visualization on Stanford2D3D dataset. (a) RGB panoramic images; (b) 360SFUDA [7]; (c) 360SFUDA++; and (d) Ground Truth (GT).

TABLE II
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.

Method	SF	mIoU	Road	S.W.	Build.	Wall	Fence	Pole	Tr.L.	Tr.S.	Veget.	Terr.	Sky	Pers.	Car	Δ
PVT [63] 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 [63] w/ 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	-
Trans4PASS [9] w/ SSL	✗	43.17	73.72	43.31	79.88	19.29	16.07	20.02	8.83	1.72	67.84	31.06	86.05	44.77	68.58	-
Trans4PASS [9] w/ MPA	✗	45.29	67.28	43.48	83.18	22.02	21.98	22.72	7.86	1.52	73.12	40.65	91.36	42.69	70.87	-
DATR-M [18] w/ CFA	✗	51.04	77.62	49.03	84.58	28.15	39.70	30.34	23.83	15.95	78.23	24.74	93.73	44.14	74.45	-
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 [11]	✓	38.21	68.78	30.71	80.37	5.26	18.95	20.90	5.25	2.36	70.19	23.30	90.20	22.55	57.90	+2.40
ProDA w/ seg-b1 [14]	✓	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 [15]	✓	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 [16]	✓	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 [13]	✓	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 [17]	✓	37.94	68.47	29.51	79.62	6.78	19.20	19.48	2.31	1.33	68.85	26.55	89.30	22.35	59.49	+2.13
SFUDA [43]	✓	41.68	72.05	37.27	81.88	14.11	24.96	23.73	0.00	0.39	70.55	33.19	90.51	29.99	63.17	+5.87
360SFUDA [7] w/ b1	✓	41.78	70.17	33.24	81.66	13.06	23.40	23.37	7.63	3.59	71.04	25.46	89.33	36.60	64.60	+5.97
360SFUDA [7] w/ b2	✓	42.18	69.99	32.28	81.34	10.62	24.35	24.29	**9.19**	**3.63**	71.28	30.04	88.75	**37.49**	65.05	+6.37
360SFUDA++ w/ b1	✓	42.53	69.65	33.49	81.90	11.35	24.80	24.90	8.92	2.14	71.79	26.58	**90.65**	35.49	64.32	+6.72
360SFUDA++ w/ b2	✓	**43.43**	**71.06**	**34.95**	**82.61**	14.23	**27.82**	**25.05**	9.18	2.82	**72.81**	**32.34**	90.39	36.20	**65.13**	**+7.62**

designated as the test set for evaluation purposes. In the context of C-to-D scenario, we employ the DensePASS training set as the unlabeled target panoramic data.

(2) SynPASS-to-DensePASS **(S-to-D):** SynPASS [9] emerges as a synthetic dataset composed of 9080 synthetic panoramic images meticulously annotated across 22 categories. Its official sets for training, validation, and testing encompass 5700, 1690, and 1690 images, respectively. For training and testing purposes, we concentrate on the subset of 13 categories that SynPASS and DensePASS datasets have in common for both training and evaluation purposes. This setup allows us to explore the nuances of transferring knowledge from synthetic (SynPASS) to real-world data (DensePASS), treating SynPASS as the synthetic source dataset and DensePASS as the real-world target dataset.

(3) Stanford2D3D-pinhole-to-panoramic **(SPin-to-SPan):** The Stanford2D3D Pinhole (SPin) dataset comprises 70496 pinhole images annotated across 13 semantic categories. Conversely, the Stanford2D3D Panoramic (SPan) dataset consists of 1413 indoor panoramic images. These images are preciesely

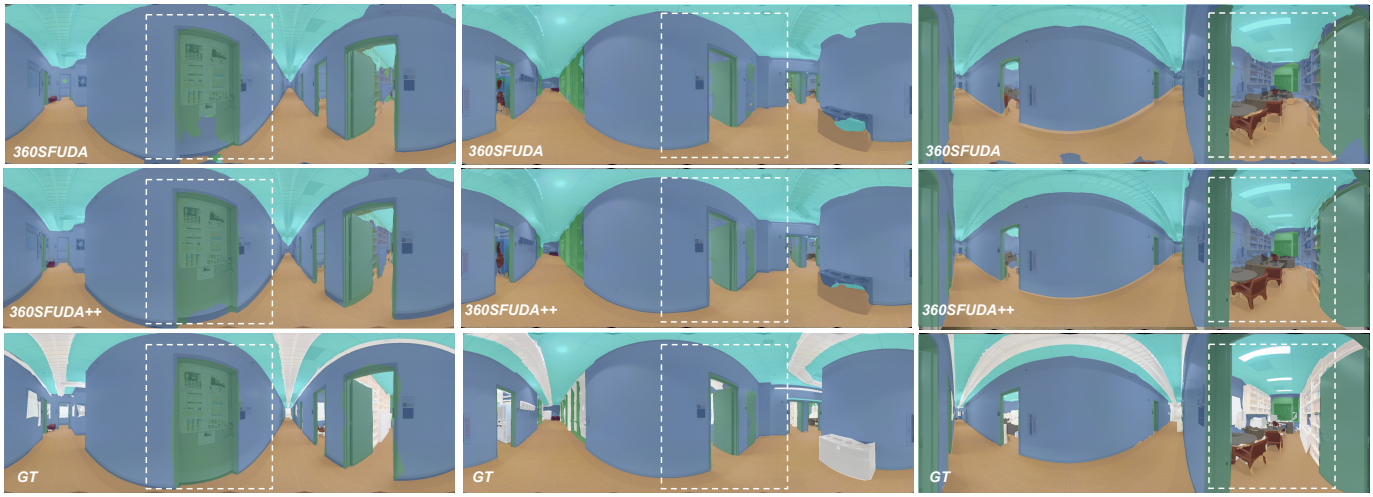

Fig. 7. Visualization on Stanford2D3D dataset. (a) RGB panoramic images; (b) 360SFUDA [7]; (c) 360SFUDA++; and (d) Ground Truth (GT).

TABLE III
EXPERIMENTAL RESULTS ON INDOOR STANFORD2D3D [64]. THE **BOLD** DENOTES THE BEST PERFORMANCE AMONG UDA AND SFUDA METHODS.

Method	SF	mIoU	Ceiling	Chair	Door	Floor	Sofa	Table	Wall	Window	Δ
PVT-S [4]	✗	57.71	85.69	51.71	18.54	90.78	34.76	65.34	74.87	39.98	-
PVT-S w/ MPA [4]	✗	57.95	85.85	51.76	18.39	90.78	35.93	65.43	75.00	40.43	-
Trans4PASS w/ MPA [4]	✗	64.52	85.08	58.72	34.97	91.12	46.25	71.72	77.58	50.75	-
Trans4PASS+ [9]	✗	63.73	90.63	62.30	24.79	92.62	35.73	73.16	78.74	51.78	-
Trans4PASS+ w/ MPA [9]	✗	67.16	90.04	64.04	42.89	91.74	38.34	71.45	81.24	57.54	-
SFDA [11]	✓	54.76	79.44	33.20	52.09	67.36	22.54	53.64	69.38	60.46	-
360SFUDA [7] w/ b1	✓	57.63	73.81	29.98	63.65	73.49	31.76	49.25	72.89	66.22	+2.87
360SFUDA [7] w/ b2	✓	65.75	82.88	38.00	**65.81**	86.71	36.32	66.10	**80.29**	69.88	+10.99
360SFUDA++ w/ b1	✓	66.54	**86.28**	**58.25**	36.04	**87.99**	48.74	65.44	77.82	**71.78**	+11.78
360SFUDA++ w/ b2	✓	**68.84**	85.50	57.59	53.15	87.40	**53.63**	**66.49**	80.23	66.75	**+14.08**

annotated with identical 13 semantic categories as observed in the pinhole dataset [64]. In this paper, our focus narrows to the 8 semantic categories that are common across both subsets.

Implementation Details. We train our frameworks with 4 NVIDIA GPUs with an initial learning rate of $6e^{-5}$, which is scheduled by the poly strategy with power 0.9 over 200 epochs. The optimizer is AdamW with epsilon $1e^{-8}$, weight decay $1e^{-4}$, and batch size is 4 on each GPU. The adaptation frameworks are trained within 10K iterations. With respect to input resolutions, training images from indoor pinhole cameras and panoramic views are processed at resolutions of 1080 $\times$ 1080 and 1024 $\times$ 512, respectively. In the SynPASS-to-DensePASS (S-to-D) scenario, synthetic panoramic images maintain a resolution of 1024 $\times$ 512. For outdoor scenarios, pinhole and synthetic panoramic images are adjusted to 1024 $\times$ 512, whereas real-world panoramic images are presented at 2048 $\times$ 400. The resolutions for validation image sets in indoor and outdoor environments are set to 2048 $\times$ 1024 and 2048 $\times$ 400, respectively, to ensure uniformity in the evaluation framework.

B. Experimental Results.

Initially, we undertake a comprehensive evaluation of the proposed 360SFUDA++ framework within the C-to-D scenario. Outcomes, including both qualitative and quantitative analyses, are presented in Fig. 5 and Tab. I, demonstrating the framework's adaptation capabilities across urban panoramas. 360SFUDA++ demonstrates superior performance relative to the previous version 360SFUDA and other source-free UDA methods [11], [13]. Remarkably, it also achieves panoramic semantic segmentation performance that are comparable to those of UDA methods that leverage source domain imagery during the adaptation process, such as Trans4PASS [9] (52.99% vs. 53.18%). Our 360SFUDA++ model yields substantial improvements, with mIoU increases of +1.41% and +2.87% using SegFormer-B1 and -B2 backbones, surpassing our previous version, 360SFUDA, and significantly outperforming SFUDA methods SFDA [11] and DATC [13] by +7.49% and +7.13% in mIoU, respectively. The qualitative results are shown in the Fig. 5, vividly illustrating our 360SFUDA++'s superior capability in handling complex segmentation tasks. We also provide the TSNE visualization to further show the superiority of our 360SFUDA++ in Fig. 8, apparently, our 360SFUDA++ gains a significant improvement in distinguishing the pixels in panoramic images in both prediction and the high-level representation spaces.

Subsequently, we conduct an evaluation of our 360SFUDA++ framework within the S-to-D scenario. The outcomes of this assessment are detailed in Tab. II and Fig. 6. Apparently, our 360SFUDA++ significantly

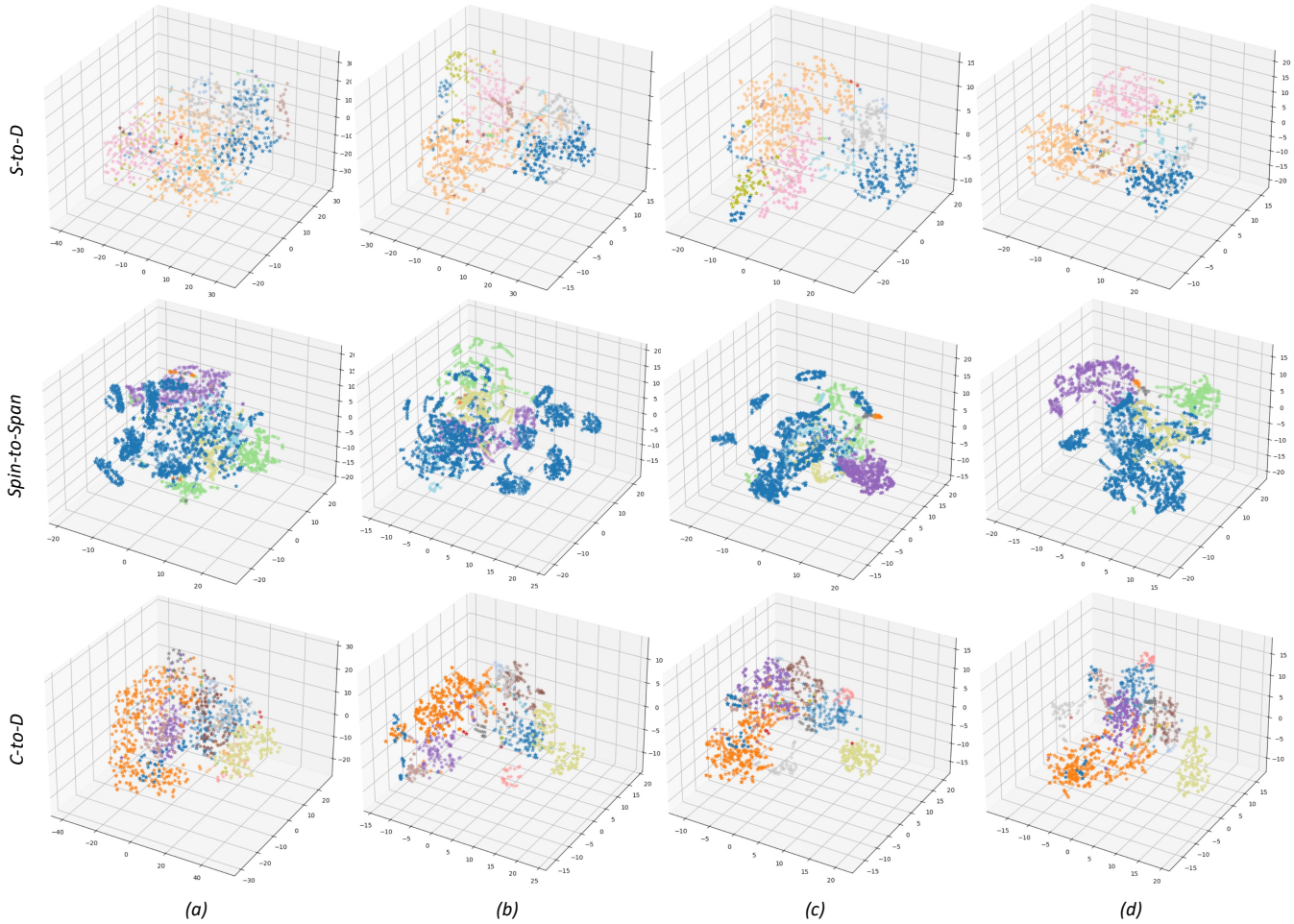

Fig. 8. TSNE Visualization on C-to-D and Spin-to-Span scenarios. (a) source, (b) SFDA [11], (c) 360SFUDA [7], and (d) 360SFUDA++.

outperforms the source-free UDA methods [11], [13] by a large margin and 360SFUDA++ improves the previous version [7] by +0.75 and +1.25 mIoU with SegFormer-B1 and -B2 backbones, respectively. The results affirm that our 360SFUDA++ framework, augmented by RP^{2}AM and CDAM, is particularly adept at facilitating robust and efficacious knowledge transfer across domains for panoramic semantic segmentation. Moreover, the performance of 360SFUDA++ in key categories pertinent to driving—such as 'person' (54.60% mIoU), 'rider' (30.76% mIoU), and 'car' (79.41% mIoU)—exemplifies its superior capability in semantic segmentation within critical traffic-related contexts. Additionally, the t-SNE visualization presented in Figure 8 further elucidates the effectiveness of 360SFUDA++ in cultivating a distinctive and informative representation space, underscoring its overall proficiency in domain adaptation for panoramic imaging tasks.

Lastly, we evaluate our 360SFUDA++ on the Spin-to-Span scenario can compare it with our previous 360SFUDA [7] and SFDA [11] methods, as well as the UDA methods, such as MPA [9]. As illustrated in Tab. III-D, 360SFUDA++ significantly outperforms the source-free method SFDA [11] by +11.78% mIoU. Notably, 360SFUDA++ even outperforms the UDA methods using source data in the adaptation (68.84% vs. 67.16% mIoU). The visualization results in Fig. 7 also demonstrate the superiority of our 360SFUDA++. Additionally, t-SNE visualizations highlight 360SFUDA++'s enhanced capability for distinguishing features within high-level representation spaces, reaffirming its effectiveness in domain adaptation for panoramic semantic segmentation. ***More quantitative comparison and visualization results refer to the supplementary material.***

TABLE IV
ABLATION STUDY OF DIFFERENT LOSS COMBINATIONS.

Loss Combinations						C-to-D		S-to-D	
$\mathcal{L}_{sup}$	$\mathcal{L}_{un}$	$\mathcal{L}_{ppa}$	$\mathcal{L}_{sft}$	$\mathcal{L}_{cda}$	$\mathcal{L}_{bns}$	mIoU	Δ	mIoU	Δ
✓						38.65	-	35.81	-
✓	✓					43.05	+4.40	36.44	+0.63
✓	✓	✓				51.05	+12.40	40.96	+5.15
✓	✓	✓	✓			51.16	+12.51	41.80	+5.99
✓				✓		44.24	+5.59	38.38	+2.57
✓				✓	✓	44.79	+6.14	38.52	+2.71
✓	✓	✓	✓	✓	✓	52.99	+14.34	42.53	+6.72

V. ABLATION STUDY

A. Different Loss Function Combinations.

Tab. IV evaluates the influences of all the proposed loss functions in 360SFUDA++, including $\mathcal{L}_{sup}$, $\mathcal{L}_{un}$, $\mathcal{L}_{ppa}$, $\mathcal{L}_{sft}$,

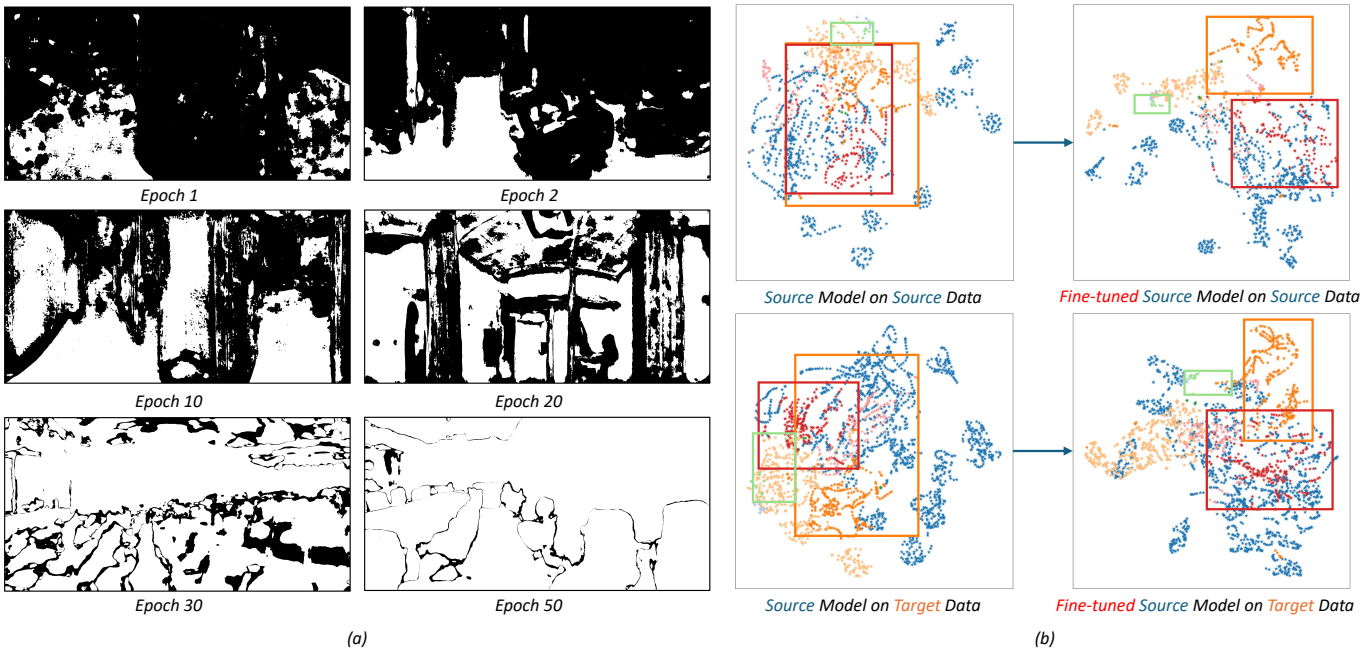

Fig. 9. (a) Confidence maps across the training procedure, white pixels represent the confident predictions while black pixels show the uncertain predictions. (b) TSNE visualization of fine-tuning source model with $\mathcal{L}_{sft}$.

$\mathcal{L}_{cda}$, and $\mathcal{L}_{bns}$. Building upon the foundational loss functions introduced in [7], 360SFUDA++ innovatively integrates $\mathcal{L}un$ to facilitate local prototypical knowledge transfer from reliable to uncertain pixels within target panoramic imagery. Moreover, it refines $\mathcal{L}ppa$ through the application of MSE between the global panoramic prototype and the confident prototype extracted from ERP images. As delineated in Tab. IV, all the proposed loss functions have a positive impact on improving the overall panoramic semantic segmentation performance. Noteworthy is the RP^{2}AM, leveraging $\mathcal{L}un$ and $\mathcal{L}ppa$, which secures substantial improvements of +12.40% and +5.15% in mIoU, underscoring the module's efficacy in optimizing knowledge extraction and adaptation processes. Furthermore, the CDAM exerts a favorable influence on segmentation outcomes, evidencing increments of +5.59% and +2.57% mIoU across C-to-D and S-to-D scenarios, respectively. These results affirm CDAM's success in replicating spatial and channel-level feature distributions between ERP and FFP images, effectively mitigating stylistic disparities across domains.

B. Ablation Study of RP^{2}AM

PPAM vs. RP^{2}AM. In order to ascertain the efficacy of the proposed RP^{2}AM in comparison to PPAM delineated in our previous work [7], we conduct ablation experiments and provide the visualization of loss curves of PPAM and RP^{2}AM on the C-to-D scenario. The results are shown in Fig. 10. The graphical representation clearly illustrates that, although RP^{2}AM's loss trajectory experiences more pronounced fluctuations during the initial stages of training, it eventually achieves a more favorable convergence compared to PPAM throughout the training duration. This observation underscores RP^{2}AM's superior capacity for stable and effective learning, marking a significant advancement over PPAM.

Effectiveness of the Confidence Maps To substantiate the efficacy of the cross-projection pixel-wise prediction assessment employed within the RP^{2}AM, we present visualizations of confidence maps throughout the training process, as illustrated in Fig. 9 (a). These visualizations reveal a progressive increase in the prevalence of white pixels within the confidence maps, symbolizing a growth in the number of pixels deemed confident. Concurrently, quantitative analyses, as detailed in Tab. V, corroborate that the implementation of reliable prototypical adaptation in 360SFUDA++ leads to considerable performance enhancements over the framework's preceding iteration detailed in [7]. This convergence of qualitative and quantitative evidence firmly establishes the beneficial impact of employing confidence maps for improved pixel-wise prediction accuracy in panoramic semantic segmentation tasks.

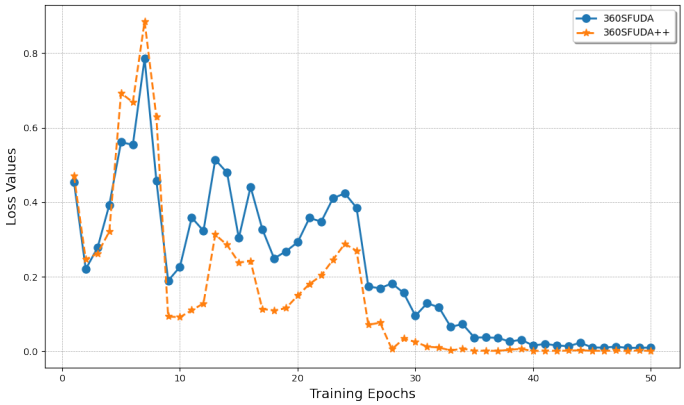

Fig. 10. Loss curves of the prototypical adaptation methods on C-to-D scenario, *i.e.*, PPAM and RP^{2}AM, respectively.

TABLE V
ABLATION OF RP^{2}AM VS. PPAM ON C-TO-D SCENARIO.

Combinations	PPAM ($\mathcal{L}_{ppa}$ in [7])	RP^{2}AM ($\mathcal{L}_{un}$+$\mathcal{L}_{ppa}$)
mIoU	45.42	**51.05**

TABLE VI
ABLATION STUDY OF THE FOV OF OUR PROPOSED FFP.

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

TABLE VII
ABLATION STUDY OF THE TRADE-OFF PARAMETERS γ FOR $\mathcal{L}_{cda}$.

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

C. Ablation Study of CDAM

Dual Attention vs. Cross Dual Attention. The dual attention (DA) approach proposed in SFDA [11] 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 projections of the same spherical data, specifically ERP and FFP, resulting in more robust and stable knowledge transfer. Moreover, in our CDA, we obtain spatial and channel characteristics across features from different projections, whereas DA operates within features. We compare DA and our cross-projection DA on the C-to-D scenario, our method achieves 44.24% mIoU, while DA only reaches 41.53% mIoU. This empirical evidence suggests that our cross-projection DA offers a more adept solution for addressing the challenges inherent in SFUDA for panoramic semantic segmentation tasks.

FoV Selection of FFP. The integration of distortion-aware mechanisms and innovative projection techniques has been extensively investigated within the scope of panoramic semantic segmentation in existing literature, such as [4], [9] and Zheng et al. [8]. Nonetheless, as delineated in Sec. I, the complex semantic content and object relationships inherent in 360-degree imagery precipitate a pronounced semantic incongruity across different domains. Intuitively, we introduce the Fixed FoV Projection (FFP) strategy to address the aforementioned problem. The quantitative results in Tab. VI show that the FoV of 90° achieves the best panoramic segmentation performance of 44.28% mIoU. This finding underscores the effectiveness of the FFP strategy in mitigating domain-specific semantic discrepancies in panoramic domain.

Ablation of γ. We study the sensitivity of hyper-parameter γ, which are the trade-off weights for the KL loss in CDAM. The quantitative results are provided in Tab. VII.

VI. DISCUSSION

Fine-tuning the Source Model. In the context of domain adaptation, the process of transferring knowledge from a pre-trained source model to a target domain presents notable challenges. The intrinsic discrepancies between the source pinhole domain and the target panoramic image domain necessitate a tailored approach for model adaptation. To this end, we advocate for the refinement of the pre-trained source model utilizing a specialized fine-tuning loss, denoted as $\mathcal{L}_{sft}$. This is predicated on the observation that a model pre-trained in the source domain may not inherently exhibit optimal performance within the target domain.

The results presented in Tab. IV substantiates the efficacy of the proposed fine-tuning loss $\mathcal{L}_{sft}$ in enhancing model performance. Further insight is afforded by the TSNE visualizations depicted in Fig. 9 (b), which contrast the performance of the source model before and after fine-tuning on both source and target domain data. These visualizations compellingly illustrate that fine-tuning with unlabeled target data not only elevates the model's effectiveness within the target domain but also augments its capacity for pixel-wise feature differentiation in the source domain. This nuanced approach underscores the importance of fine-tuning pre-trained models with domain-specific adaptations to bridge the gap between disparate domains, thereby improving model generalizability and performance across different modalities.

Difference between Pinhole-to-Panoramic and Pinhole-to-Pinhole UDA. In the context of SFUDA, the primary challenge arises from the domain shift attributable to style discrepancies, which stem from differing data distribution between the source and target domains. However, when the objective is to adapt models from the conventional pinhole camera domain to the domain of panoramic images, one encounters additional complexities. As evidenced by the results in Tba. I, endeavors have been made for adaptation within the pinhole camera domain [11], [13] tend to offer marginal improvements in performance when bechmarked against a standard baseline. This phenomenon is largely due to the more intricate nature of domain shift that occur in the transition from pinhole to panoramic adaptation scenarios, a topic elaborated upon in Sec. I. By integrating our proposed modules, *i.e.*, RP^{2}AM and CDAM, the performance has been largely improved by +6.08% and +5.72% performance gain over SFDA [11] and DATC [13], respectively.

VII. CONCLUSION AND FUTURE WORK

In this paper, we explored the challenging source-free unsupervised domain adaptation from pinhole to panoramic image domain. To this end, we proposed a novel framework for panoramic semantic segmentation, namely 360SFUDA++, to address the domain shifts between pinhole and panoramic images, including semantic mismatch, inevitable distortion, and inherent style discrepancies. We subtly utilized the multi-projection versatility of 360° data for efficient and reliable domain knowledge transfer. The RP^{2}AM is introduced to select the confident extracted knowledge and integrate panoramic prototypes for reliable knowledge adaptation. Meanwhile, we proposed CDAM to better align the spatial and channel characteristics across projections at the feature level between domains. Extensive experiments on the synthetic and real-world benchmarks, including indoor and outdoor scenarios, show that our 360SFUDA++ outperforms prior approaches and is on par with the UDA methods using the source data.

Future Work: In future research, we intend to employ Large Language Models (LLMs) and Multi-modal Large Language Models (MLLMs) to address the domain discrepancies, specifically targeting semantic mismatches between pinhole and

panoramic images. This approach aims to enhance the interpretative alignment and integration across the different projection of spherical data. Also, the Segment Anything Model (SAM) can be applied to unleash its powerful zero-shot instance segmentation ability.

REFERENCES

[1] H. Ai, Z. Cao, J. Zhu, H. Bai, Y. Chen, and L. Wang, "Deep learning for omnidirectional vision: A survey and new perspectives," *arXiv preprint arXiv:2205.10468*, 2022. 1

[2] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang, "Pass: Panoramic annular semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4171–4185, 2019. 1

[3] K. Yang, X. Hu, Y. Fang, K. Wang, and R. Stiefelhagen, "Omnisupervised omnidirectional semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1

[4] J. Zhang, K. Yang, C. Ma, S. Reiß, K. Peng, and R. 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*, 2022, pp. 16 917–16 927. 1, 2, 3, 4, 7, 9, 12

[5] J. Zhang, C. Ma, K. Yang, A. Roitberg, K. Peng, and R. Stiefelhagen, "Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation," *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1, 7

[6] X. Yue, Z. Zheng, S. Zhang, Y. Gao, T. Darrell, K. Keutzer, and A. S. 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*, 2021, pp. 13 834–13 844. 1, 7

[7] X. Zheng, P. Zhou, A. Vasilakos, and L. Wang, "Semantics, distortion, and style matter: Towards source-free uda for panoramic segmentation," 2024. 1, 2, 3, 4, 7, 8, 9, 10, 11

[8] X. Zheng, J. Zhu, Y. Liu, Z. Cao, C. Fu, and L. Wang, "Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation," *arXiv preprint arXiv:2303.14360*, 2023. 1, 2, 3, 4, 5, 7, 12

[9] J. Zhang, K. Yang, H. Shi, S. Reiß, K. Peng, C. Ma, H. Fu, K. Wang, and R. 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, 3, 4, 8, 9, 10, 12

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 1, 2

[11] Y. Liu, W. Zhang, and J. Wang, "Source-free domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1215–1224. 2, 3, 4, 6, 7, 8, 9, 10, 12

[12] M. Ye, J. Zhang, J. Ouyang, and D. Yuan, "Source data-free unsupervised domain adaptation for semantic segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2233–2242. 2, 3, 4

[13] C.-Y. Yang, Y.-J. Kuo, and C.-T. 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)*. IEEE, 2022, pp. 1–6. 2, 3, 4, 7, 8, 9, 10, 12

[14] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. 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*, 2021, pp. 12 414–12 424. 2, 8

[15] J. N. Kundu, A. Kulkarni, A. Singh, V. Jampani, and R. V. Babu, "Generalize then adapt: Source-free domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7046–7056. 2, 3, 8

[16] J. Huang, D. Guan, A. Xiao, and S. Lu, "Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3635–3649, 2021. 2, 3, 8

[17] X. Guo, J. Liu, T. Liu, and Y. Yuan, "Simt: Handling open-set noise for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7032–7041. 2, 8

[18] X. Zheng, T. Pan, Y. Luo, and L. 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*, 2023, pp. 18 687–18 698. 2, 4, 7, 8

[19] W. Zhang, Y. Liu, X. Zheng, and L. Wang, "Goodsam: Bridging domain and capacity gaps via segment anything model for distortion-aware panoramic semantic segmentation," *arXiv preprint arXiv:2403.16370*, 2024. 3

[20] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," *Advances in neural information processing systems*, vol. 32, 2019. 3

[21] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2090–2099. 3

[22] L. Hoyer, D. Dai, and L. 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*, 2022, pp. 9924–9935. 3, 7

[23] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Dada: Depth-aware domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7364–7373. 3

[24] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305. 3

[25] S. Stan and M. Rostami, "Unsupervised model adaptation for continual semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2593–2601. 3

[26] F. Fleuret *et al.*, "Uncertainty reduction for model adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9613–9623. 3

[27] W. Shen, Q. Wang, H. Jiang, S. Li, and J. Yin, "Unsupervised domain adaptation for semantic segmentation via self-supervision," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2747–2750. 3

[28] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526. 3

[29] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3764–3773. 3

[30] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 384–15 394. 3

[31] X. Zheng, C. Fu, H. Xie, J. Chen, X. Wang, and C.-W. Sham, "Uncertainty-aware deep co-training for semi-supervised medical image segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106051, 2022. 3

[32] J. Chen, C. Fu, H. Xie, X. Zheng, R. Geng, and C.-W. Sham, "Uncertainty teacher with dense focal loss for semi-supervised medical image segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106034, 2022. 3

[33] J. Zhu, Y. Luo, X. Zheng, H. Wang, and L. 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*, 2023, pp. 11 720–11 730. 3

[34] X. Zheng, Y. Luo, H. Wang, C. Fu, and L. Wang, "Transformer-cnn cohort: Semi-supervised semantic segmentation by the best of both students," *arXiv preprint arXiv:2209.02178*, 2022. 3

[35] J. Chen, D. Deguchi, C. Zhang, X. Zheng, and H. Murase, "Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation," *Available at SSRN 4617170*. 3

[36] X. Zheng, Y. Luo, P. Zhou, and L. Wang, "Distilling efficient vision transformers from cnns for semantic segmentation," *arXiv preprint arXiv:2310.07265*, 2023. 3

[37] J. Chen, D. Deguchi, C. Zhang, X. Zheng, and H. Murase, "Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation," *arXiv preprint arXiv:2310.02296*, 2023. 3

[38] H. Xie, C. Fu, X. Zheng, Y. Zheng, C.-W. Sham, and X. Wang, "Adversarial co-training for semantic segmentation over medical images," *Computers in biology and medicine*, vol. 157, p. 106736, 2023. 3

[39] X. Zheng and L. Wang, "Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition," *arXiv preprint arXiv:2403.14082*, 2024. 3

[40] H.-W. Yeh, B. Yang, P. C. Yuen, and T. Harada, "Sofa: Source-data-free feature alignment for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 474–483. 3

[41] Y. Zhao, Z. Zhong, Z. Luo, G. H. Lee, and N. Sebe, "Source-free open compound domain adaptation in semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7019–7032, 2022. 3

[42] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. B. Ayed, "Source-free domain adaptation for image segmentation," *Medical Image Analysis*, vol. 82, p. 102617, 2022. 3

[43] Y. Tian, J. Li, H. Fu, L. Zhu, L. Yu, and L. Wan, "Self-mining the confident prototypes for source-free unsupervised domain adaptation in image segmentation," *IEEE Transactions on Multimedia*, 2024. 3, 7, 8

[44] Z. Sun, L. Lin, and Y. Yu, "You only label once: A self-adaptive clustering-based method for source-free active domain adaptation," *IET Image Processing*, 2024. 3

[45] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *ICML*, 2018. 3

[46] J. Choi, T. Kim, and C. Kim, "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6829–6839, 2019. 3

[47] S. Sankaranarayanan, Y. Balaji, A. Jain, S.-N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3752–3761, 2018. 3

[48] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7472–7481, 2018. 3

[49] M. Liu, S. Wang, Y. Guo, Y. He, and H. Xue, "Pano-sfmlearner: Self-supervised multi-task learning of depth and semantics in panoramic videos," *IEEE Signal Processing Letters*, vol. 28, pp. 832–836, 2021. 3

[50] C. Zhang, Z. Cui, C. Chen, S. Liu, B. Zeng, H. Bao, and Y. Zhang, "Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12 612–12 621, 2021. 3

[51] Q. Wang, D. Dai, L. Hoyer, O. Fink, and L. V. Gool, "Domain adaptive semantic segmentation with self-supervised depth estimation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8495–8505, 2021. 3

[52] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2039–2049, 2017. 3

[53] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6929–6938, 2019. 3

[54] Z. Murez, S. Kolouri, D. J. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4500–4509, 2018. 3

[55] C. Chen, W. Xie, T. Xu, W. Huang, Y. Rong, X. Ding, Y. Huang, and J. Huang, "Progressive feature alignment for unsupervised domain adaptation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 627–636, 2019. 3

[56] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *ArXiv*, vol. abs/1612.02649, 2016. 3

[57] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. 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)*, pp. 2502–2511, 2019. 3

[58] L. Melas-Kyriazi and A. K. Manrai, "Pixmatch: Unsupervised domain adaptation via pixelwise consistency training," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 430–12 440, 2021. 3

[59] M. Eder, M. Shvets, J. Lim, and J.-M. Frahm, "Tangent images for mitigating spherical distortion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 426–12 434. 4

[60] Y. Li, Y. Guo, Z. Yan, X. Huang, Y. Duan, and L. Ren, "Omnifusion: 360 monocular depth estimation via geometry-aware fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2801–2810. 4

[61] K. Yang, J. Zhang, S. Reiß, X. Hu, and R. Stiefelhagen, "Capturing omni-range context for omnidirectional segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1376–1386. 7

[62] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 635–12 644. 7

[63] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578. 8

[64] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017. 9

[65] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[66] C. Ma, J. Zhang, K. Yang, A. Roitberg, and R. Stiefelhagen, "Densepass: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2766–2772. 7