# Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation

Xu Zheng[2] Jinjing Zhu[1] Yexin Liu[1] Zidong Cao[1] Chong Fu[2,4] Lin Wang[1,3*]
[1]AI Thrust, HKUST(GZ) [2]Northeastern University [3]Dept. of CSE, HKUST
[4]Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, NEU, China

zhengxu128@gmail.com, zhujinjing.hkust@gmail.com, yliu292@connect.hkust-gz.edu.cn
caozidong1996@gmail.com, fuchong@mail.neu.edu.cn, linwang@ust.hk

## Abstract

*The ability of scene understanding has sparked active research for panoramic image semantic segmentation. However, the performance is hampered by distortion of the equirectangular projection (ERP) and a lack of pixel-wise annotations. For this reason, some works treat the ERP and pinhole images equally and transfer knowledge from the pinhole to ERP images via unsupervised domain adaptation (UDA). However, they fail to handle the domain gaps caused by: 1) the inherent differences between camera sensors and captured scenes; 2) the distinct image formats (e.g., ERP and pinhole images). In this paper, we propose a novel yet flexible dual-path UDA framework, DPPASS, taking ERP and tangent projection (TP) images as inputs. To reduce the domain gaps, we propose cross-projection and intra-projection training. The cross-projection training includes tangent-wise feature contrastive training and prediction consistency training. That is, the former formulates the features with the same projection locations as positive examples and vice versa, for the models' awareness of distortion, while the latter ensures the consistency of cross-model predictions between the ERP and TP. Moreover, adversarial intra-projection training is proposed to reduce the inherent gap, between the features of the pinhole images and those of the ERP and TP images, respectively. Importantly, the TP path can be freely removed after training, leading to no additional inference cost. Extensive experiments on two benchmarks show that our DPPASS achieves +1.06% mIoU increment than the state-of-the-art approaches.* https://vlis2022.github.io/cvpr23/DPPASS

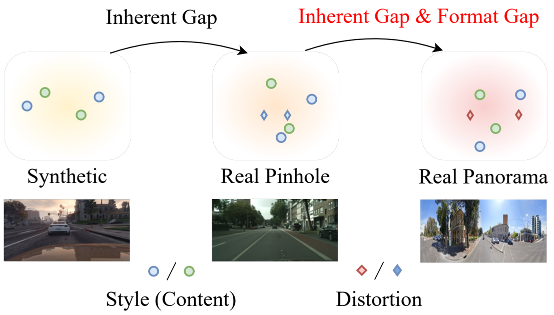

Figure 1. We tackle a new problem by addressing two types of domain gaps, *i.e*., the inherent gap (style) and format gap (distortion) between the pinhole and panoramic (360°) images.

## 1. Introduction

Increasing attention has been paid to the emerging 360° cameras for their omnidirectional scene perception abilities with a broader field of view (FoV) than the traditional pinhole images [1]. Intuitively, the ability to understand the surrounding environment from the panoramic images has triggered the research for semantic segmentation as it is pivotal to practical applications, such as autonomous driving [45, 50] and augmented reality [28]. Equirectangular projection (ERP) [46] is the most commonly used projection type for the 360° images [1] and can provide a complete view of the scene. However, the ERP type suffers from severe distortion in the polar regions, resulting in noticeable object deformation. This significantly degrades the performance of the pixel-wise dense prediction tasks, *e.g*., semantic segmentation. Some attempts have been made to design the convolution filters for feature extraction [32,50,53]; however, the specifically designed networks are less generalizable to other spherical image data. Moreover, labeled datasets are scarce, thus making it difficult to train effective 360° image segmentation models.

To tackle these issues, some methods, *e.g*., [50] treat the ERP and pinhole images equally, like the basic UDA task,

*Corresponding author.

[1] Here, panoramic and 360° images are interchangeably used.

and directly alleviate the mismatch between ERP and pinhole images by adapting the neural networks trained in the pinhole domain to the 360° domain via unsupervised domain adaptation (UDA). For instance, DensePASS [23] proposes a generic framework based on different variants of attention-augmented modules. Though these methods can relieve the need for the annotated 360° image data [50], they fail to handle the existing domain gaps caused by: 1) diverse camera sensors and captured scenes; 2) distinct image representation formats (ERP and pinhole images) and yield unsatisfied segmentation performance. Accordingly, we define these two types of domain gaps as the inherent gap and format gap (See Fig. 1).

In this paper, we consider using the tangent projection (TP) along with the ERP. It has been shown that TP, the geometric projection [7] of the 360° data, suffers from less distortion than the ERP. Moreover, the deep neural network (DNN) models designed for the pinhole images can be directly applied [10]. To this end, we propose a novel dual-path UDA framework, dubbed DPPASS, taking ERP and TP images as inputs to each path. The reason is that the ERP provides a holistic view while TP provides a patch-wise view of a given scene. For this, the pinhole images (source domain) are also transformed to the pseudo ERP and TP formats as inputs. To the best of our knowledge, our work takes the first effort to leverage two projection formats, ERP and TP, to tackle the inherent and format gaps for panoramic image semantic segmentation. Importantly, the TP path can be freely removed after training, therefore, no extra inference cost is induced.

Specifically, as shown in Fig. 2, the cross-projection training is proposed at both the feature and prediction levels for tackling the challenging format gap (Sec. 3.2). At the feature level, the tangent-wise feature contrastive training aims at mimicking the tangent-wise features with the same distortion and discerning the features with distinct distortion, to further learn distortion-aware models and decrease the format gap. Meanwhile, the less distorted tangent images are used in the prediction consistency training. It ensures the consistency between the TP predictions and the tangent projections of the ERP predictions for models' awareness of the distortion variations. For the long-existing inherent gap, the intra-projection training imposes the style and content similarities between the features from the source and target domains for both the ERP and TP images (Sec. 3.3). As such, we can reduce the large inherent and format gaps between the 360° and pinhole images by taking advantage of dual projections.

We conduct extensive experiments from the pinhole dataset, Cityscapes [6], to two 360° datasets: DensePASS [23] and WildPASS [44]. The experimental results show that our framework surpasses the existing SOTA methods by 1.06% on the DensePASS test set. In summary, our main contributions are summarized as follows: (I) We study a new problem by re-defining the domain gaps between 360° images and pinhole images as two types: the inherent gap and format gap. (II) We propose the first UDA framework taking ERP and tangent images to reduce the types of domain gaps for semantic segmentation. (III) We propose corss- and intra- projection training that take the ERP and TP at the prediction and feature levels to reduce the domain gaps.

## 2. Related work

**Panoramic image semantic segmentation** Most existing works [22,39,55] on semantic segmentation focused on pinhole images having a limited FoV; consequently, the performance significantly drops when they are applied to the 360° images. The panoramic image segmentation has to tackle two challenges: 1) the inevitable distortion and object deformation in the ERP images and 2) a lack of accurately labeled data [8, 24, 36, 38].

In the literature, the mainstream methods can be divided into three types: supervised learning methods [30, 41], UDA methods [13, 23, 49], and unsupervised contrastive learning methods [18]. Among the supervised learning methods, [19, 32, 48] focus on designing distortion-aware and trainable deformable convolution layers for dense depth and semantic prediction on the panoramic images. [41, 42, 44], on the other hand, explore the multi-source learning schemes to train the network on the pinhole images and deploy it to the unseen panoramas. The unsupervised contrastive learning methods learn robust feature representations, allowing the network model to better generalize to data from a different distribution [18].

As labeled panoramic image data is limited, UDA methods [23,49] have been proposed to transfer knowledge from the output, and feature space of pinhole images to those of the panoramic images. However, these methods do not fully consider the domain gaps between the pinhole and 360° images. Also, the domain gap between different projection types, *e.g.*, ERP and TP, of 360° images has been neglected. For this reason, we formulate the two types of domain gaps between the 360° and pinhole images: 1) inherent gaps caused by the different sensors and scenes and 2) format gap caused by the difference of image representation formats. In our work, we are the first to tackle both domain gaps simultaneously.

**Unsupervised domain adaptation.** UDA takes the labeled source domain data and unlabeled target domain data as inputs and trains the network model in an unsupervised manner to enhance the generalization capacity to the targeted domain. The UDA techniques are critical for semantic segmentation, especially when labeled data is particularly scarce. The mainstream UDA methods rely on self-training with pseudo labels [14, 51], adversarial learn-

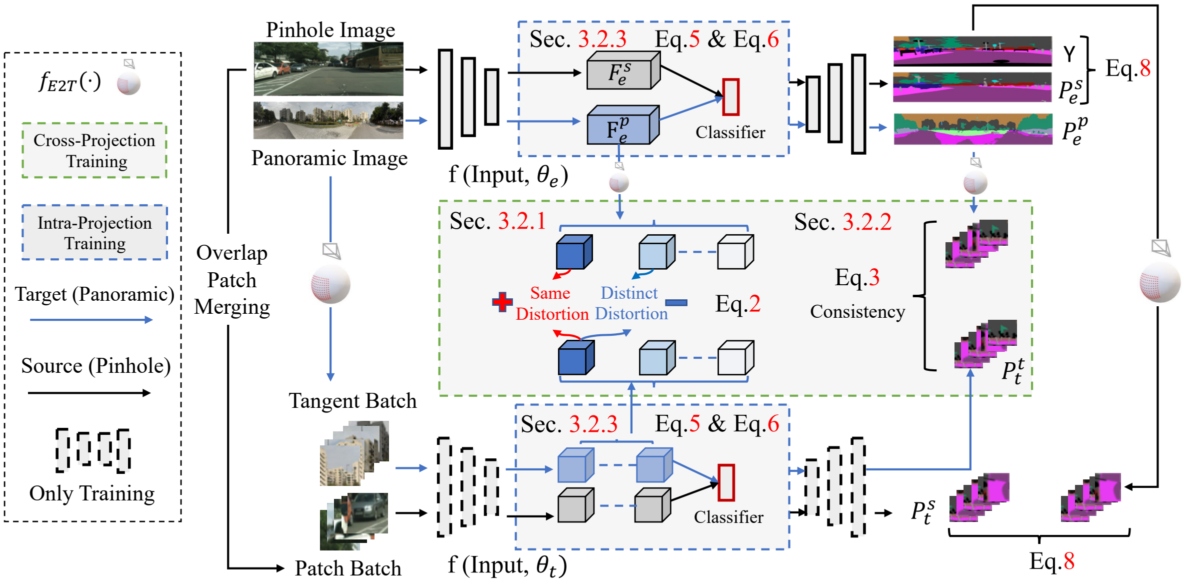

Figure 2. Overview of the proposed DPPASS framework, consisting of two models $f(Input, \theta_e)$, $f(Input, \theta_t)$. Our method has two major components: cross- and intra- projection training. The cross-projection training explores prediction consistency training and tangent-wise feature contrastive training, and the intra-projection training employs adversarial training for transferring knowledge from the pinhole images to panoramas.

ing [2, 5, 15, 34], entropy minimization [9, 33], and self-ensembling [11, 17, 26]. The self-training methods create pseudo labels for the target data and gradually adapt through the iterative improvement [31]. However, pseudo labels are often error-prone. To mitigate this problem, [29, 35, 56] attempt to make self-training less sensitive to the incorrect pseudo labels. The adversarial learning methods mainly adopt generative adversarial networks (GANs) to learn a shared latent representation between the two domains while maintaining the domain-specific characteristics. The entropy minimization methods aim to enforce structural consistency across domains by applying it jointly with the square losses [4] or adversarial loss [37]. Our work explores the potential of adversarial learning and ensemble learning to tackle the two types of domain gaps between the pinhole and 360° images, mentioned in the introduction. We focus on exploiting the feature embeddings and predictions from the ERP and TP paths to transfer knowledge from the pinhole image domain to the 360° image domain.

# 3. Methodology

## 3.1. Overview

An overview of the proposed DPPASS is depicted in Fig. 2. Given the target domain data consisting a set of $n$ unlabeled ERP $P$ = $E_p^1$, ..., $E_p^n$ and a set of annotated pinhole images in the source domain are transformed to the $m$ pseudo ERP $S$ = ($I_s^1$, $Y_s^1$), ..., ($I_s^m$, $Y_s^m$). The tangent images $T$ = $E_t^1$, ..., $E_t^{18n}$ are projected by function $f_{E2T}(\cdot)$ (see Fig. 4) from the ERP image set $P$, and the pseudo tangent image set $T^*$ = $I_{t*}^1$, ..., $I_{t*}^{18m}$ are projected by the overlap patch merging from the pseudo ERP set $S$. $E_p^i$ is the $i$-th ERP image with spatial dimensions $H \times W$, $I_s^i$ is the $i$-th pseudo ERP (pinhole) image from the source domain has the same spatial dimensions as $E_p^i$ with its corresponding pixel-level label $Y_s^i \in (1, C)^{H \times W}$, where $C$ is the number of classes. $I_t^i$ and $I_{t*}^i$ are the tangent and pseudo tangent images projected from $E_p^i$ and $I_s^i$. We propose a novel dual-projection UDA framework to minimize the two domain gaps, namely the inherent gap and format gap, simultaneously for panoramic image semantic segmentation. The two network models $f(Input, \theta_e)$ and $f(Input, \theta_t)$ in the framework are based on the vision transformers [40]. $f(Input, \theta_e)$ takes an ERP image $E_p^i \in P$ and the pseudo ERP $I_s^i \in S$ as the input, and $f(Input, \theta_t)$ takes the tangent images $I_t^i$ and the pseudo tangent images $I_{t_*}^i$ as inputs:

$$\begin{aligned} P_e^p, F_e^p = f(E_p^i, \theta_e), P_e^s, F_e^s = f(I_s^i, \theta_e), \\ P_t^t, F_t^t = f(E_t^i, \theta_t), P_t^s, F_t^s = f(I_{t_*}^i, \theta_t), \end{aligned} \tag{1}$$

where $P_e^p$ and $P_e^s$ are the ERP format predictions, $P_t^t$ and $P_t^s$ are the tangent format predictions, $F_e^p$, $F_e^s$, $F_t^t$ and $F_t^s$ are the high-level features. The model $f(Input, \theta_e)$ in the ERP path and the model $f(Input, \theta_t)$ in the TP path are trained and optimized individually. Importantly, the TP path $f(Input, \theta_t)$ can be removed after training, and only $f(Input, \theta_e)$ is used for inference, leading to no additional inference cost.

Based on the aforementioned definitions for the two types of domain gaps, our proposed framework consists of two key components. Firstly, cross-projection training is

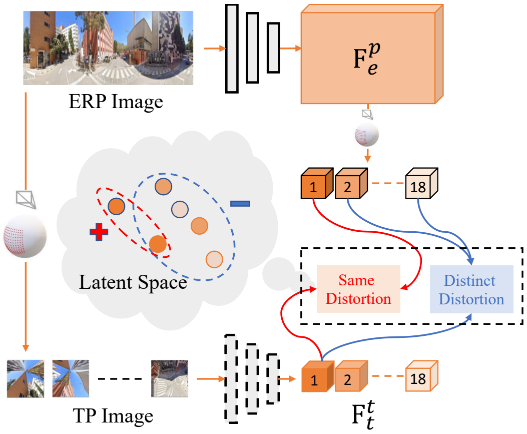

Figure 3. Overview of the proposed the tangent-wise feature contrastive training (TFCT) module.

proposed in both the prediction and feature spaces. As the network models in two paths exploit ERP and TP images as the representations of the 360° data, $f(P, \theta_e)$ and $f(T, \theta_t)$ can learn the inherent gap and format gap together from the patch-wise view of TP images and the holistic view of ERP images based on the feature and prediction perspectives. Secondly, the intra-projection training is designed in each path to tackle the inherent domain gap.

In the cross-projection training, the prediction consistency training is proposed to fully utilize the distinct characteristics of tangent images to reduce the format gap. Because the tangent images have less distortion and object deformation, we first leverage the geometric projection function $f_{E2T}(\cdot)$ to transform the ERP image to the TP patches. Then, the consistency regularization loss is employed to ensure the consistency of the cross-model predictions. For the feature-level cross-projection training, tangent-wise feature contrastive training (TFCT) is proposed to align the tangent-wise high-level features to reduce the format gap. We now describe these components in detail

## 3.2. Cross-Projection Training

### 3.2.1 Tangent-wise feature contrastive training

We explore to impose consistency between the ERP and TP paths at the feature map level to diminish the format gap. As the models in the ERP path and TP path take the ERP and tangent images as inputs, respectively, the extracted high-level features are heterogeneous to each other (see Fig. 3). To align the features from the dual projection paths, we propose to process the features $F_e^p$ from $f(P, \theta_e)$ by $f_{E2T}(\cdot)$ to match the features $F_t^t$ extracted from $f(T, \theta_t)$. After aligning the features $F_t^t = F_{t1}^t, F_{t2}^t, ..., F_{t18}^t$ and $f_{E2T}(F_e^p) = F_{e1}^p, F_{e2}^p, ..., F_{e18}^p$ from two models, a contrastive learning strategy is applied to reduce the format gaps caused by the distortion. This is because tangent images are the geometric projection of the 360° image data; thus the distortion at different locations is different.

For this reason, as shown in Fig. 3, we divide the features according to the TP $f_{E2T}(\cdot)$, and then align them by the projection locations. Specifically, the TP images, which are oriented in the same position and angle, have the same distortion and object deformation. That is, in Fig. 3, for the feature representation (cube), only the ones in the same color having the same projection locations are the positive example pairs, and the other representations in this batch are all negative example pairs. Consequently, our proposed TFCT module aims to maximize the consistency between the positive pairs, *e.g.*, $F_{t1}^t$ and $F_{e1}^p$, which represent the same form of distortion of the TP format. By contrast, the tangent features extracted from tangent images at different projection locations, are formulated as negative pairs, *e.g.*, $F_{t1}^t$ and $F_{e2}^p$. Given the two feature sequences $F_t^t = F_{t1}^t, F_{t2}^t, ..., F_{t18}^t$ and $f_{E2T}(F_e^p) = F_{e1}^p, F_{e2}^p, ..., F_{e18}^p$, we formulate the contrastive training loss, based on the InfoNCE [25], which is:

$$L_{fc} = \frac{1}{F_i} \sum_{f_+ \in F_i} -log \frac{exp(f_+/\tau)}{exp(f_+/\tau) + \sum_f exp(f_-/\tau)}, \quad (2)$$

where $f_+$ denotes the positive examples which stand for the same position in ERP, *e.g.*, $F_{t1}^t$ and $F_{e1}^p$, the negative examples $f_-$ denote the tangent-wise features extracted from different locations in ERP, *e.g.*, $F_{t1}^t$ and $F_{e2}^p$, $F_i$ denotes all the tangent-wise features in one batch and the $\tau$ is the temperature hyper-parameter.

Previous methods *e.g.*, [18], applying contrastive learning to panoramic semantic segmentation, have to maintain a large memory bank to store the negative examples. This leads to a large capacity of memory and high training costs. Our TFCT module takes the tangent images projected from distinct locations in the same training batch as the negative examples; therefore, it requires less memory during the training process.

### 3.2.2 Prediction consistency training

We present the prediction consistency training to address the format gap which is mainly caused by the distortion of the ERP images. ERP is a common spherical image representation format based on the simple relation between rectangular and spherical coordinates, making it suffer from severe image distortion and object deformation. These inevitable problems impede applying UDA to the pinhole and panoramic image domains.

Recently, TP [10] is shown to better mitigate the distortion, and thus the deep learning models developed for the pinhole images can be directly applied to the TP images. Though TP has less distortion than the ERP, ERP has a more holistic awareness of the surrounding scene. Accordingly, we leverage ERP and tangent images together to bridge the domain gaps caused by the distortion. Specifically, as depicted in Fig. 4, the network model $f(Input, \theta_e)$ in the ERP path predicts a semantic label from an ERP image input while the model $f(Input, \theta_t)$ in the TP path processes the tangent image patches. We then project the ERP prediction $P_e^p$ with $f_{E2T}(\cdot)$ to get patch prediction maps $f_{E2T}(P_e^p)$ of the TP format. The consistency regularization is finally applied between $P_t^t$ and $f_{E2T}(P_e^p)$ to make the network models in the dual paths aware of the distortion discrepancies between the ERP and TP images. For convenience, the consistency regularization loss is formulated by the KL-Divergence:

$$\mathcal{L}_{pc} = \sum_{i=1}^{18} f_{E2T}(P_{ei}^p) \log \frac{f_{E2T}(P_{ei}^p))}{P_{ti}^t}, \tag{3}$$

where $P_{ei}^p$ and $P_{ti}^t$ denote the $i$th tangent-wise predictions. The distinct representations of the same sphere data have the same semantic and content information, thus leading to better dual models' awareness of the distortion differences between ERP and TP. The tangent projection also can be treated as a data augmentation approach, which is thus applied before and after the model forward prediction for the consistency training.

### 3.3. Intra-Projection Training

We propose intra-projection training to decrease the inherent domain gap between panoramic images and pinhole images in each projection path. The main goal is to regularize the learning of the internal features from source and target domains, which are extracted from the same model and in the same format (*i.e.*, ERP and TP). Specifically, a domain classifier $f(F, \theta_d)$ is added after the feature extractor to distinguish the features extracted from panoramas (TP images) or pinhole images (pseudo TP images).

We denote $d$ as the binary variable for the extracted features $F$, which indicates whether $F$ is extracted from the panoramas ($F_e^p$ and $F_e^s$ if $d_t$ = 1) or from the pinhole images ($F_t^t$ and $F_t^s$ if $d_s$ = 0). The classifier and the feature extractor (encoder) are optimized individually, the classifier is trained to better distinguish the features extracted from different domains while the feature extractor is optimized to generate domain-invariant features:

$$D_u = f(F_e^p, \theta_d), D_s = f(F_e^s, \theta_d), \tag{4}$$

where the $D_t$ and $D_s$ are the classification predictions of the features. For training the classifier, we aim to minimize the supervised loss with Binary Cross Entropy (BCE) as:

$$\begin{aligned} L_d^c = -[&(d_t \cdot log(D_t) + (1-d_t) \cdot log(1-D_t)) \\ &+(d_s \cdot log(D_s) + (1-d_s) \cdot log(1-D_s))]. \end{aligned} \tag{5}$$

To reduce the inherent domain gap, the domain adaptation loss for the feature extractor (encoder) is measured by BCE as follows:

$$\begin{aligned} L_d = -[&(d_t \cdot log(D_s) + (1-d_t) \cdot log(1-D_s)) \\ &+(d_s \cdot log(D_t) + (1-d_s) \cdot log(1-D_t))]. \end{aligned} \tag{6}$$

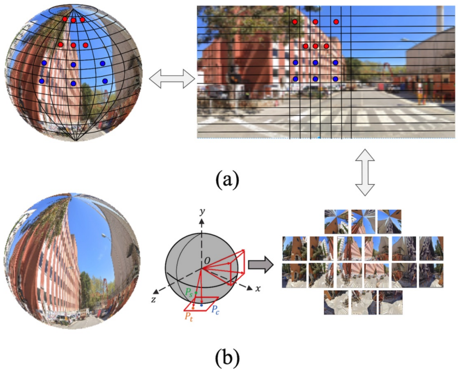

Figure 4. (a) Visualization of distortion. (b) Tangent image facilitates the transferable and scalable panoramic image representation. We use 18 tangent images to project the ERP as in [21].

In summary, through the dual-projection regularization, we make the two models, $f(P, \theta_e)$ and $f(T, \theta_t)$, learn the domain gaps between pinhole and panoramic images at ERP and tangent image scales.

### 3.4. Optimization

The training objective containing four losses is defined as:

$$\mathcal{L} = \mathcal{L}_s + \alpha \cdot \mathcal{L}_d + \beta \cdot \mathcal{L}_{pc} + \mathcal{L}_{fc}, \tag{7}$$

where the $L_s$ is the supervised loss on the Cityscapes dataset, the $L_g$ refers to the intra-projection loss, $L_{pc}$ denotes the prediction consistency loss, the $L_{fc}$ is the tangent-wise feature contrastive loss between $f(Input, \theta_e)$ and $f(Input, \theta_t)$ and the $\alpha$ and $\beta$ are the trade-off weight of the proposed loss terms. The supervised loss on Cityscapes is formulated using the standard Cross-Entropy (CE) loss:

$$\mathcal{L}_s = -\sum_{i=0}^{C} Y_i log(P_e^s). \tag{8}$$

Especially, for the network model in the TP path, the pseudo tangent images are obtained through random crops on the Cityscapes train set.

## 4. Experiments

To evaluate the performance of our method, we conduct extensive experiments on two benchmark datasets including DensePASS [23] and WildPASS [44]. Experimental results demonstrate the superiority of our proposed DPPASS.

### 4.1. Datasets and Implementation Details

**Cityscapes [6]** is an autonomous driving dataset that contains urban street scenes recorded from 50 different cities with precise pixel-wise annotations of 19 semantic categories. **DensePASS [23]** is a panoramic dataset and contains 2,000 images for training and 100 precise annotated images for testing. **WildPASS [44]** is a panoramic dataset

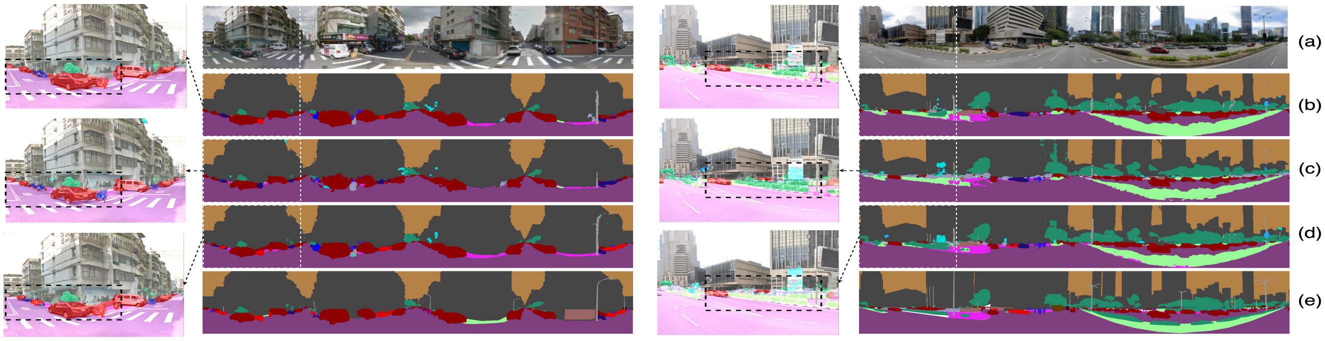

Figure 5. Example visualization results from DensePASS test set. (a) Input, (b) Fully supervised Segformer-B1 without domain adaptation [40], (c) Trans4PASS-T [50], (d) DPPASS-T, and (e) Ground truth.

| Network | Backbone | CS | DP | GAPs |
|---|---|---|---|---|
| PSPNet [52] | ResNet-50 | 78.6 | 29.5 | 49.1 |
| | ResNet-101 | 79.8 | 30.4 | 49.4 |
| DeepLabv3+ [3] | ResNet-50 | 80.1 | 29.0 | 51.1 |
| | ResNet-101 | 80.9 | 32.5 | 48.4 |
| Semantic-FPN [20] | ResNet-50 | 74.5 | 29.9 | 44.6 |
| | ResNet-101 | 75.8 | 28.8 | 47.0 |
| | PVT-T | 71.5 | 31.2 | 40.3 |
| SETR [54] | Transformer-L | 77.9 | 36.1 | 41.8 |
| Segformer [40] | Mit-B1 | 78.5 | 38.5 | 40.0 |
| | MiT-B2 | 81.0 | 42.4 | 38.6 |
| ERFNet [27] | ERFNet | 72.1 | 16.7 | 55.4 |
| FANet [16] | ResNet-34 | 71.3 | 26.9 | 44.4 |
| DANet [12] | ResNet-50 | 79.3 | 28.5 | 50.8 |
| P2PDA [49] | ERFNet | 72.1 | 34.1 | 38.0 |
| | ResNet-34 | 71.3 | 33.1 | 38.2 |
| | ResNet-50 | 79.3 | 39.8 | 39.5 |
| Tarns4PASS [50] | Trans4PASS-T | 79.1 | 41.5 | 37.6 |
| | Trans4PASS-S | 81.1 | 44.8 | 36.3 |
| DPPASS(Ours) | ResNet-34 | 75.4 | 38.9 | 36.5 |
| | ResNet-50 | 78.6 | 42.3 | 36.3 |
| | Mit-B1 | 76.3 | 42.4 | 36.1 |
| | Mit-B2 | 80.1 | **48.6** | **32.4** |

Table 1. Performance gaps of semantic segmentation methods from Cityscapes dataset (CS) to DensePASS dataset (DP).

designed to capture diverse scenes from all around the globe and contains 2500 panoramas.

**Evaluation.** We take the mean Intersection-over-Union (mIoU) as the evaluation metric in both the source and target domains. Our framework is evaluated on the DensePASS / WildPASS validation set via test at a single scale, and the resolution is 400 × 2048.

**Implementation details.** Our framework is implemented with Pytorch and trained on multiple NVIDIA GPUs. Both models in our framework are based on the efficient Segformer [40]. Ours-T and Ours-S are two implementations of our framework, which are based on SeformerB1 and SegormerB2, respectively.

### 4.2. Inevitable Domain Gaps

As shown in Tab. 1, there are large segmentation performance drops from Cityscapes to DensePASS datasets. Even though some recent high-performance transformer-based networks [40, 54] have better results on the pinhole images (Cityscapes) than the convolutional neural networks (CNNs), the performance on the panoramic images is still unsatisfying. Meanwhile, although some well-designed distortion-aware frameworks, *e.g.*, [50], have been proposed for panoramic semantic segmentation, the domain gaps and performance drops are still large.

Our proposed DPPASS, empowered by the unified vision transformer backbone MiT [40] without using the deformable components, achieves better performance than the SOTA methods. The reported results of our DPPASS on DensePASS are the average predictions of the two models. With Mit-B2 backbone, our method achieves 48.6% mIoU on DensePASS test set and the performance drop is only 32.4%, which is 3.9% mIoU better than the SOTA method Trans4PASS (32.4% vs. 36.3%). This indicates that our method effectively reduces the inherent gap and format gap by exploring the knowledge from ERP and TP paths, thus yielding better segmentation performance.

### 4.3. Experimental Results

We first train our DPPASS with the DenPASS train set and Cityscapes train set and then evaluate with the DensePASS test set. Tab. 2 shows the quantitative results. We compare our framework with some SOTA panoramic segmentation methods: PASS [41], Omni-sup [43] and Trans4PASS [50]; domain adaptation approaches: P2PDA [49] and PCS [47]. Without the well-designed distortion-aware components, like De-

| Method | mIoU | road | sidewalk | building | wall | fense | pole | traffic Light | traffic Sign | tegetation | terrain | sky | Person | rider | car | truck | bus | train | motorcycle | bicycle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ERFNet | 16.65 | 63.59 | 18.22 | 47.01 | 9.45 | 12.79 | 17.00 | 8.12 | 6.41 | 34.24 | 10.15 | 18.43 | 4.96 | 2.31 | 46.03 | 3.19 | 0.59 | 0.00 | 8.30 | 5.55 |
| PASS(ERFNet) | 23.66 | 67.84 | 28.75 | 59.69 | 19.96 | 29.41 | 8.26 | 4.54 | 8.07 | 64.96 | 13.75 | 33.50 | 12.87 | 3.17 | 48.26 | 2.17 | 0.82 | 0.29 | 23.76 | 19.46 |
| Omni-sup(ECANet) | 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(Adversarial) | 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 |
| PCS | 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 |
| Trans4PASS-T † | 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 |
| Trans4PASS-S † | 55.22 | 78.38 | 41.58 | 86.48 | 31.54 | 45.54 | 33.92 | 22.96 | 18.27 | 79.40 | 41.07 | 93.82 | 48.85 | 23.36 | 81.02 | 67.31 | 69.53 | 86.13 | 60.85 | 39.09 |
| DPPASS-T(Ours) | **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 |
| DPPASS-S(Ours) | **56.28** | 78.99 | **48.14** | **87.63** | 42.12 | 44.85 | 34.95 | **27.38** | **19.21** | 78.55 | 43.08 | 92.83 | **55.99** | 29.10 | 80.95 | 61.42 | 55.68 | 79.70 | **70.42** | 38.40 |

Table 2. Per-class results of the SOTA panoramic image semantic segmentation methods on DensePASS test set.

| Method | Backbone | mIoU(%) |
|---|---|---|
| Source domain Supervised | Segformer-B1<br>Segformer-B2 | 47.90<br>54.11 |
| Trans4PASS-T [50] | Segformer-B1 | 54.67 |
| Trans4PASS-S [50] | Segformer-B2 | 62.91 |
| DPPASS-T(Ours) | Segformer-B1 | **60.38** |
| DPPASS-S(Ours) | Segformer-B2 | **63.53** |

Table 3. Experimental results of the SOTA panoramic image semantic segmentation methods on WildPASS test set.

formable Patch Embedding (DPE) and Deformable MLP (DMLP) [50], our DPPASS-T and DPPASS-S using the unified Segformer outperform the SOTA segmentation method, Trans4PASS-T, and Trans4PASS-S by 2.12% and 1.06%, respectively. Meanwhile, compared with the SOTA UDA segmentation methods, our DPPASS also yields the best performance. Specifically, our DPPASS achieves 1.47% and 2.45% mIoU increment with Segformer-B1 and Segformer-B2 backbones than the UDA method PCS [47]. This indicates our DPPASS better tackles two types of domain gaps (*i.e*., inherent and format gaps) between 360° and pinhole image domain.

Fig. 5 shows the qualitative comparison with the supervised Segformer-1 [40], Trans4PASS [50] and our DM-PASS. For panoramic images, the larger objects have more complex distortion than the smaller ones, thus it is more difficult to completely and neatly segment these large objects, such as sidewalks, walls, etc. Obviously, our DPPASS has *significantly better* segmentation results on these larger objects with greater distortion, as shown in Fig. 5. The quantitative results in Tab. 2 also show that our DPPASS-S outperforms the Trans4PASS-S by 6.56% IoU on the sidewalk class. For the classes for autonomous driving, such as persons, riders, and motorcycles, the white dotted boxes in Fig. 5 show that our DPPASS-S achieves much better segmentation results than the Trans4PASS-S. The IoU increments of person, rider, and motorcycle categories are +7.14%, +5.74%, and +9.57%, respectively.

We also evaluate our DM-PASS on the WildPASS test set. Tab. 3 shows the quantitative results, where our DPPASS approach consistently outperforms the SOTA UDA segmentation method Trans4PASS [50]. The mIoU improvements of our DPPASS-T and DPPASS-S over the Trans4PASS-T and Trans4PASS-S are 5.71% and 0.62%, respectively. It is worth noting that our DPPASS achieves a dramatic increase of mIoU by 5.71% higher than the SOTA method Trans4PASS (60.38% vs. 54.67%) with the Segformor-B1 backbone. This validates the problem that naively treating the ERP and pinhole images equally leads to less optimal UDA segmentation performance. Therefore, the two types of domain gaps between the 360° and pinhole image domain, defined by our work, are pivotal.

## 5. Ablation study and Analysis

**Dual Projection vs. Single Projection.** The dual-projection training means ERP and tangent images are individually used in dual paths while the single projection only uses the ERP in both paths. So the results of dual projection (49.53%) and single projection (45.22%) show the superiority of dual-projection training. This also indicates the tangent projection can bring complementary benefits to the standard domain alignment.

**Rationality of Prediction Consistency Training** The tangent projection of the same sphere data utilized in our work can be formulated as a data augment operation. In classic consistency training procedure, data augment methods are always leveraged before and after the forward propagation process. In our work, we use the tangent projection before and after the forward propagation to make our model aware of the distortion variation. Numerically, with our prediction consistency training, +7.13% mIoU improvement is obtained than the source pretrained model.

**Rationality of Feature Contrastive Training** Since the tangent projection $f_{E2T}(\cdot)$ is performed on the ERP input and the ERP features equally without shuffling the location, the two sequences of features are in one-to-one correspondence. Intuitively, we leverage this correspondence to distinguish the distortion information. As shown in Tab. 4, the feature-wise contrastive learning give +4.90% mIoU increment compared with the supervised baseline. Qualitatively,

| Losses | | | | mIoU | $\triangle$ |
|---|---|---|---|---|---|
| $\mathcal{L}_s$ | $\mathcal{L}_g$ | $\mathcal{L}_{pc}$ | $\mathcal{L}_{fc}$ | | |
| ✓ | | | | 42.40 | - |
| ✓ | ✓ | | | 50.12 | +7.72 |
| ✓ | | ✓ | | 49.53 | +7.13 |
| ✓ | | | ✓ | 47.30 | +4.90 |
| ✓ | ✓ | ✓ | ✓ | 55.30 | +12.9 |

Table 4. Ablation study of different loss combinations with DS-PASS-T framework on the DensePASS test set.

| | Tangent Projection | | | | |
|---|---|---|---|---|---|
| Size | 96 × 96 | 144 × 144 | 224 × 224 | 384 × 384 | 512 × 512 |
| mIoU | 49.98 | 52.22 | **55.30** | 55.17 | 52.56 |

Table 5. Ablation study of different Tangent-Projection size with DS-PASS-T framework on the DensePASS test set.

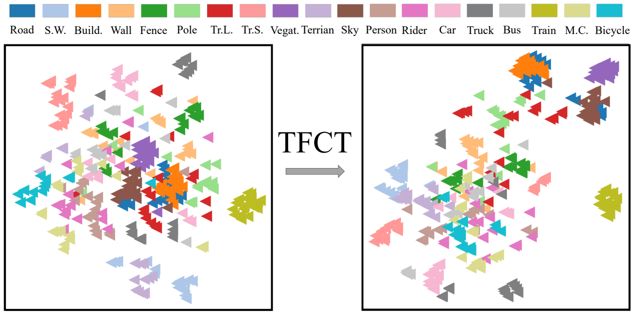

Figure 6. TSNE visualization of features with and without our proposed TFCT module.

the feature embeddings are shown in Fig. 6. As our TFCT acts on the representation space and provides feature-wise alignment, the features are pushed closer to each category against distortion and object deformation.

**Rationality of Intra-Projection Training.** Both ERP and TP images provide critical domain knowledge, including style and distortion features. Intuitively, we propose to utilize this critical information in both projections to facilitate knowledge transfer. The adversarial module in each path makes the models learn complementary domain knowledge in different scales (whole ERP & Tangent Patch). Numerically, as shown in Tab. 4, the Intra-Projection loss $\mathcal{L}_g$ achieves +7.72% mIoU increment than the baseline.

| $\alpha$ | 0.001 | 0.01 | 0.02 | 0.05 | 0.1 |
|---|---|---|---|---|---|
| mIoU | 48.25 | 50.01 | **50.12** | 48.12 | 47.44 |
| $\beta$ | 10 | 20 | 50 | 100 | 200 |
| mIoU | 46.97 | 47.21 | **49.53** | 49.38 | 42.1 |

Table 6. Ablation study of hyper-parameters $\alpha$ and $\beta$. The reported results are trained with the combination of the supervised loss and the loss terms $L_{pc}$ and $L_{fc}$ based on our DPPASS-T.

**Loss functions.** We conduct ablation experiments on the DensePASS dataset to analyze the impact of the supervised loss $\mathcal{L}_s$, the intra-porjection loss $\mathcal{L}_g$ (Eq. 6), the prediction consistency training loss $\mathcal{L}_{pc}$ (Eq. 3) and the TFCT loss $\mathcal{L}_{fc}$ (Eq. 2) in DPPASS. In Tab. 4, different combinations of losses are applied. It is obvious that the intra-porjection loss $\mathcal{L}_g$ reduces the inherent gaps which are caused by the different sensors, and brings an increase of 4.93% in the mIoU. For the cross-model modules, we can see that the prediction consistency training loss $\mathcal{L}_{pc}$ which imposes the prediction consistency between different representations of the same spherical data gives an improvement of 7.13% in mIoU. As for the TFCT $\mathcal{L}_{fc}$, it also contributes positively to the decrease of the large domain gaps between the 360° image domain and the pinhole image domain well with 4.90% mIoU improvement over supervised baseline.

**TP patch size**. Tab. 5 reports the mIoU(%) with different TP patch sizes on the DensePASS dataset. It shows that the optimal patch size is 224 × 224. Too large or too small projection size impedes the segmentation performance, the best trade-off patch size is 224 × 224.

**Trade-off ratio of $\alpha$ and $\beta$**. Tab. 6 reports the mIoU(%) of our DPPASS-T with different ratios of $\alpha$ and $\beta$ on the DensePASS test set. The best trade-off weights for $\mathcal{L}_{pc}$ and $\mathcal{L}_{fc}$ are $\alpha$ = 0.02 $\beta$ = 50.

**Inference cost**. For one ERP with the size of 400 × 2048, the inference costs of our DPPASS-S and the Trans4PASS-S are 108.84G and 251.08G in FLOPs, respectively. Our DPPASS reduce 60% computational cost while achieving +1% increment than the SOTA Trans4PASS method.

## 6. Conclusion

In this paper, we studied a new problem by refining the domain gaps between the panoramic and pinhole images into two types: the inherent gap and the format gap. We accordingly proposed DPPASS, the first dual-projection UDA framework, taking ERP and tangent images as input to each path to reduce the domain gaps. We introduced intra-projection training to reduce the inherent gap while the format gap was addressed by the cross-projection training. Importantly, the TP path can be removed after training, adding no extra inference cost. Our DPPASS significantly surpassed the prior UDA methods for panoramic image semantic segmentation and achieved new SOTA performance.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45.

## 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] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019. 3

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6

[4] 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. 3

[5] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9082–9091, 2021. 3

[6] 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. 2, 5

[7] Harold Scott Macdonald Coxeter. Introduction to geometry. 1961. 2

[8] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4350–4362, 2019. 2

[9] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 982–991, 2019. 3

[10] 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. 2, 4

[11] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017. 3

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 6

[13] Qiqi Gu, Qianyu Zhou, Minghao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Pit: Position-invariant transform for cross-fov domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8761–8770, 2021. 2

[14] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2021. 2

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[16] Ping Hu, Federico Perazzi, Fabian Caba Heilbron, Oliver Wang, Zhe Lin, Kate Saenko, and Stan Sclaroff. Real-time semantic segmentation with fast attention. *IEEE Robotics and Automation Letters*, 6(1):263–270, 2020. 6

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 3

[18] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1421–1427. IEEE, 2021. 2, 4

[19] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. *arXiv preprint arXiv:1901.02039*, 2019. 2

[20] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 6

[21] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Duan Ye, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *2022 Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 2022. 5

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 2

[23] 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. 2, 5

[24] Kenichi Narioka, Hiroki Nishimura, Takayuki Itamochi, and Teppei Inomata. Understanding 3d semantic structure around the vehicle with monocular cameras. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 132–137. IEEE, 2018. 2

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[26] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. 3

[27] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017. 6

[28] Ronald Schroeter. Inception of perception—augmented reality in virtual reality: Prototyping human–machine interfaces for automated driving. In *User Experience Design in the Era of Automated Driving*, pages 477–503. Springer, 2022. 1

[29] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *European conference on computer vision*, pages 532–548. Springer, 2020. 3

[30] Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9442–9451, 2019. 2

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[32] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018. 1, 2

[33] Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1358–1368, 2021. 3

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[35] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 3

[36] Robert Varga, Arthur Costea, Horatiu Florea, Ion Giosan, and Sergiu Nedevschi. Super-sensor for 360-degree environment perception: Point cloud segmentation using image features. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017. 2

[37] 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. 3

[38] Chunxiang Wang, Hengrun Zhang, Ming Yang, Xudong Wang, Lei Ye, and Chunzhao Guo. Automatic parking based on a bird's eye view vision system. *Advances in Mechanical Engineering*, 6:847406, 2014. 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 548–558. IEEE, 2021. 2

[40] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12077–12090, 2021. 3, 6, 7

[41] 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. 2, 6

[42] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelhagen. Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 457–464. IEEE, 2020. 2

[43] Kailun Yang, Xinxin Hu, Yicheng Fang, Kaiwei Wang, and Rainer Stiefelhagen. Omnisupervised omnidirectional semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 6

[44] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30:1866–1881, 2021. 2, 5

[45] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1376–1386, 2021. 1

[46] Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5677–5686, 2022. 1

[47] 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. 6, 7

[48] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3533–3541, 2019. 2

[49] 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. 2, 6

[50] 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, 6, 7

[51] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 2020–2030, 2017. 2

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 6

[53] Qiang Zhao, Chen Zhu, Feng Dai, Yike Ma, Guoqing Jin, and Yongdong Zhang. Distortion-aware cnns for spherical images. In *IJCAI*, pages 1198–1204, 2018. 1

[54] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6881–6890. Computer Vision Foundation / IEEE, 2021. 6

[55] 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. 2

[56] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 3