Bending Reality: Distortion-aware Transformers for Adapting to Panoramic Semantic Segmentation

Jiaming Zhang\textsuperscript{1} Kailun Yang\textsuperscript{1,*} Chaoxiang Ma\textsuperscript{2} Simon Reiß\textsuperscript{1,3} Kunyu Peng\textsuperscript{1} Rainer Stiefelhagen\textsuperscript{1}
\textsuperscript{1}CV:HCI Lab, Karlsruhe Institute of Technology \textsuperscript{2}ByteDance Inc. \textsuperscript{3}Carl Zeiss AG

Abstract

Panoramic images with their 360° directional view encompass exhaustive information about the surrounding space, providing a rich foundation for scene understanding. To unfold this potential in the form of robust panoramic segmentation models, large quantities of expensive, pixel-wise annotations are crucial for success. Such annotations are available, but predominantly for narrow-angle, pinhole-camera images which, off the shelf, serve as sub-optimal resources for training panoramic models. Distortions and the distinct image-feature distribution in 360° panoramas impede the transfer from the annotation-rich pinhole domain and therefore come with a big dent in performance. To get around this domain difference and bring together semantic annotations from pinhole- and 360° surround-visuals, we propose to learn object deformations and panoramic image distortions in the Deformable Patch Embedding (DPE) and Deformable MLP (DMLP) components which blend into our Transformer for PAnoramic Semantic Segmentation (Trans4PASS) model. Finally, we tie together shared semantics in pinhole- and panoramic feature embeddings by generating multi-scale prototype features and aligning them in our Mutual Prototypical Adaptation (MPA) for unsupervised domain adaptation. On the indoor Stanford2D3D dataset, our Trans4PASS with MPA maintains comparable performance to fully-supervised state-of-the-arts, cutting the need for over 1,400 labeled panoramas. On the outdoor DensePASS dataset, we break state-of-the-art by 14.39% mIoU and set the new bar at 56.38%.

1. Introduction

Panoramic 360° cameras have received an increasing amount of attention in fields, such as omnidirectional sensing in automated vehicles [13, 75] and bringing immersive viewing experiences to augmented- and virtual reality displays [69, 71]. Opposed to images captured with pinhole cameras, that occupy narrow Fields of View (FoV), panoramic images offer omni-range perception, benefiting the detection of road scene objects and indoor scene elements [13, 20]. In particular, dense semantic segmentation on panoramic images, facilitates a high-level holistic pixel-wise understanding of surrounding environments [45, 73].

Panoramic semantic segmentation is usually performed on 2D panoramas that were transformed using equirectangular projection [58, 75], which is accompanied by image distortions and object deformations (see Fig. 1). Further, in the 360° image domain, labeled data is scarce which necessitates model training to be carried out on semantically matching narrow-FoV pinhole datasets. These two circumstances culminate in a significantly degraded performance on panoramic segmentation as compared to the pinhole counterpart [72] and as such they have to be adequately addressed. Considering the intricacies of panoramas, convolution variants [10, 55, 59] and attention-augmented models [75] were proposed to mitigate image distortions and enlarge receptive fields of Convolutional Neural Networks (CNNs). However, they remain sub-optimal in handling the severe deformations from pinhole- to panoramic data, and fail in establishing long-range contextual dependencies in the ultra-wide 360° images, which prove essential for accurate semantic segmentation [17, 94].
In light of these challenges, we propose a Transformer for Panoramic Semantic Segmentation (Trans4PASS) architecture, and overcome image distortions and object deformations with two novel design choices: Our Deformable Patch Embedding (DPE) is located at the early image sequentialization- and intermediate feature interpretation stages empowering the model to learn characteristic panoramic image distortions and preserve semantics. Secondly, with the Deformable MLP (DMLP) module in the feature parsing stage, we mix patches with learned spatial offsets to enhance global context modeling.

The challenging mismatch between the label-rich pinhole- and the label-scarce panoramic domain can also be addressed by unsupervised domain adaptation (UDA), considering labeled 2D Pinhole images as source- and 360° Panoramas as target domain. Following previous works [45, 75], we refer to this scenario as PIN2PAN. Taking this view on the learning problem, shows to be a vital ingredient for circumventing the expensive panoramic image annotation process while satisfying the need for large-scale annotated data [94] to train robust segmentation transformers. Unlike common adversarial-learning [44] and pseudo-label self-learning [97] methods for UDA, we put forward Mutual Prototypical Adaptation (MPA), which generates mutual prototypes for pinhole- and panoramic multi-scale feature embeddings, distilling prototypical knowledge of both domains, which proves advantageous to domain-separate distillation [84]. On top, we show MPA works with pseudo-labels in a joint manner and provides a complementary alignment incentive in the feature space.

To verify the capability for generalization to diverse scenarios of our solution, we evaluate Trans4PASS on both indoor- and outdoor panoramic-view datasets, i.e., Stanford2D3D [1] and DensePASS [45] benchmarks. On DensePASS, it outperforms the previous best result [88] by >10.0% in mIoU. Our solution achieves top performance among unsupervised methods on Stanford2D3D and even ranks higher than many competing supervised methods.

In summary, we deliver the following contributions:

1. We consider panoramic deformations in our distortion-aware Transformer for Panoramic Semantic Segmentation (Trans4PASS) with deformable patch embedding- and deformable MLP modules.

2. We present Mutual Prototypical Adaptation to transfer models via distilling dual-domain prototypical knowledge, boosting performance by coupling it with pseudo-labels in feature- and output space.

3. Our framework for transferring models from PIN2PAN yields excellent results on two competitive benchmarks: On Stanford2D3D we circumvent using 1,400 expensive panorama labels while achieving comparable results and on DensePASS we boost state-of-the-art performance by an absolute 14.39% in mIoU.

2. Related Work

Semantic- and panoramic segmentation. Dense semantic segmentation is experiencing steep progress since FCN [43] addressed it end-to-end. Following works built upon FCN to improve performance by enlarging receptive fields [23, 93] and refining context priors [29, 80]. Driven by non-local blocks [66], self-attention [63] is integrated to learn long-range dependencies [17, 26] within FCNs. Currently, architectures which replace convolutional- with transformer-based backbones [15, 61] emerge. Then, image perception is viewed from the lens of sequence-to-sequence learning with dense prediction transformers [42, 82] and semantic segmentation transformers [57, 94]. Recently, MLP-like architectures [36, 39, 60] which alternate spatial- and channel mixing sparked interest for recognition tasks. Most methods are designed for narrow-FoV images and often have large accuracy drops in the 360° domain. In this work, we address panoramic segmentation, with a novel Transformer architecture which considers a broad FoV already in its design and handles the panorama-specific semantic distribution via MLP-based mixing.

By capturing wide-FoV scenes, panoramic images can serve as starting point for a more holistic scene understanding. Outdoor panorama segmentation works rely on fisheye cameras [14, 54, 77, 79] or panoramic images [27, 47, 70, 74] for seamless 360° parsing. Indoor methods on the other hand focus on either distortion-mitigated representations [28, 33, 55] or multi-task schemes [40, 58, 85]. Most of these works assume that labeled images are available in the target panorama domain. We cut this requirement for labeled target data and circumvent the prohibitively expensive annotation process of determining pixel-wise semantics in complex real-world surroundings. Therefore, unlike previous works, we look through the lens of unsupervised transfer learning and introduce a pinhole- to panorama (PIN2PAN) adaptation method to profit from rich, readily available annotated pinhole datasets. In experiments, our panoramic segmentation transformer architecture generalizes to both indoor and outdoor scenes.

Unsupervised domain adaptation. Domain adaptation has been thoroughly investigated to enhance model generalization to unseen domains, with two predominant paradigms based either on self-training [8, 21, 92] or adversarial learning [2, 22, 62]. Self-training methods generally create pseudo-labels to gradually adapt through iterative improvement [37], whereas adversarial solutions leverage the idea of GANs [18] to perform image translation [22, 35], or enforce alignment in layout matching [25, 34] and feature agreement [44, 45]. Further adaptation flavors, consider uncertainty reduction [56, 95], model ensembling [4, 76], category-level alignment [41, 46], or adversarial entropy minimization [49, 64]. Relevant to our work, PIT [19] addresses the camera gap with FoV-based adaptation, whereas
P2PDA [45] first tackles PIn2PA transfer by learning attention correspondences. Aside from distortion-adaptive architecture design, we revisit PIN2PA segmentation from a feature prototype adaptation-based perspective where we distill panoramic knowledge through class-wise prototypes. Different from methods using individual prototypes for source and target domains [84, 91], we present mutual prototypical adaptation, which jointly exploits source and target features to boost transfer beyond the FoV.

3. Methodology

Here, we put forward our panoramic semantic segmentation framework. In Sec. 3.1, we introduce the Trans4PASS architecture for capturing distortion-aware features and long-range dependencies, with detailed descriptions of deformable patch embeddings and the deformable MLP module in Sec. 3.2 and 3.3. Finally, we outline our domain adaptation method using mutual prototype features in Sec. 3.4.

3.1. Trans4PASS Architecture

To investigate the transformer model on panoramic semantic segmentation, we create two versions of Trans4PASS models (T: Tiny and S: Small). We build both with four stages, where for the tiny model, each stage encompasses 2 layers, for the small version the stages have 3, 4, 6, and 3 layers. As shown in Fig. 2, the pyramidal stages are inspired by recent transformers [65, 68], which reduce the feature scales in deeper layers. Given an input image with $H \times W \times 3$, Trans4PASS makes use of a Patch Embedding (PE) module [68] to split the image into patches. To deal with the severe distortions in panoramas, a special Deformable Patch Embedding (DPE) module is proposed and applied in the encoder and decoder (Fig. 2c).

In the encoder, each feature map $f_l \in \{f_1, f_2, f_3, f_4\}$ in the $l^{th}$ stage is down-sampled by the $l^{th}$ stride $\in \{4, 8, 16, 32\}$. The channel dimensions $C_l \in \{64, 128, 320, 512\}$ grow successively. Different from the FPN-like decoder [94] and vanilla-MLP based decoder [68] in Fig. 2, we propose the Deformable MLP (DMLP) decoder structure, which mixes feature patches extracted via DPE. Given the extracted feature hierarchy in multiple scales from the encoder, four deformable decoder layers process the feature hierarchy into a consistent shape of $\frac{H'}{2} \times \frac{W'}{2} \times C_{emb}$, where we set the number of resulting embedding channels $C_{emb}=128$. An ensuing linear layer transforms the 128 channel output to contain the number of semantic classes of the respective task.

3.2. Deformable Patch Embedding

Spherical topological images captured by 360° cameras occupy a polar coordinate system with $\theta \in [0, 2\pi]$ and $\phi \in [0, \pi]$. To represent it in 2D space, the spherical data is usually converted into a panoramic format in euclidean-like space through the equirectangular projection. This process leads to severe shape distortions in the projected panoramic image, as seen in Fig. 1. Therefore, a common PE module with fixed sampling positions does not respect these shape distortions of objects and the overall scene. Inspired by deformable convolution [12] and overlapping PE [68], we propose Deformable Patch Embeddings (DPE) and employ them on the input to the encoder and the decoder, splitting panoramic images and features. Given an input image or feature map $f \in \mathbb{R}^{H \times W \times C_{in}}$, a standard PE module [15, 68] splits it into a flattened 2D path sequence $z \in \mathbb{R}^{\frac{HW}{s^2} \times C_{in}}$, where $\frac{HW}{s^2}$ is the number of patches and $s$ is the width and height of each patch. Each element in this sequence is passed through a linear projection layer transforming it into $C_{out}$ dimensional embeddings.

Consider a single patch in $z$ representing a rectangle of size $s \times s$ with $s^2$ positions. We can define a position offset relative to a location $(i, j)[i, j] \in [1, s]$ in the patch as $\Delta_{(i, j)} \in \mathbb{N}^2$. In standard PE, these offsets are fixed and lie in $\Delta_{(i, j)} \in \{\frac{3}{2}, \frac{3}{2}\}^2$. Take e.g. a $3 \times 3$ patch, offsets $\Delta_{(i, j)}$ relative to the center will lie in $[-1, 1] \times [-1, 1]$. As we want to process panoramic images, which inherit distortions from the equirectangular projection, we can directly address this degradation in the PE. To this end, in our Deformable Patch Embedding (DPE), we enable the model to learn a data-dependent offset $\Delta^{DPE} \in \mathbb{N}^{H \times W \times 2}$ that can better cope with the spatial connections of objects, as present in distorted patches. DPE is learnable and pre-
dicts relative offsets based on the original input $f$. The offset $\Delta_{DPE}^{(i,j)}$ is calculated as depicted in Eq. (1).

$$
\Delta_{DPE}^{(i,j)} = \left[ \min(\max(-\frac{H}{W}, g(f(i,j))), \frac{H}{W}) \right] - \left[ \min(\max(-\frac{H}{W}, g(f(i,j))), \frac{H}{W}) \right],
$$

where $g(\cdot)$ is the offset prediction function, which we implement via the deformable convolution operation [12]. The hyperparameter $r$ puts a constraint onto the offsets and is set as 4 in our experiments. The learned offsets make DPE adaptive and as a result distortion-aware.

In earlier works, DPT [7] applies non-overlapping PE with anchor-based offsets at later stages, PS-ViT [83] uses a progressive sampling module coupled with previous iterations, and Deformable DETR [96] leverages deformable attention to enhance feature maps. Unlike these previous works, our proposed DPE is designed for pixel-dense prediction tasks and is flexible to replace the raw PE without having to couple previous iterations. Intuitively, a model supplied with DPE, can profit from pinhole images and better adapt to distortions in panoramic images by learning to counteract severe deformations in the data.

### 3.3. Deformable MLP

Apart from the specific design of the encoder, the decoder with an adaptive feature parsing capacity is crucial in segmentation transformers [68, 89]. As shown in Fig. 2a, some transformers [94] borrow a FPN-like decoder from the CNN counterpart [38], whose receptive field is limited to the feature resolution in its final stage [65]. SegFormer [68] takes inspiration from Multilayer Perceptron-based (MLP) models [60] and integrates a vanilla MLP to combine features (Fig. 2b), but does not consider potential distortions in the imaging data. Next, we propose a mechanism to associate self-attention in Transformers and deformation-properties in 360° imagery. Linking both of these enables profiting from long-range dependencies for dense scene parsing and keeping this improvement when processing panoramic scenes. Achieving this distortion-aware property at manageable computational complexity, we put forward the **Deformable MLP (DMLP)** module. Within each stage of the decoder, DMLP mixes patches across the channel dimension, but with a particularly large receptive field, which improves the interpretation of features delivered by the aforementioned DPE.

Fig. 3 shows the difference in MLP-based modeling: while the vanilla MLP (see Fig. 3a) performs traditional linear projection without learning any spatial context, CycleMLP (see Fig. 3b) has a limited spatial receptive field by hand-crafted, fixed offsets in mixing patches and their channels. In Fig. 3c, the proposed DMLP generates a learned spatial offset (top) in a wider range and an adaptive manner. Given the input feature map $f \in \mathbb{R}^{H \times W \times C_{in}}$, the spatial offset $\Delta_{DMLP}^{(i,j,c)}$ is predicted channel-wise as in Eq. (1) and is then flattened as $\Delta_{DMLP}^{(k,c)}$, where $k \in HW$ and $c \in C_{in}$, for mixing the flattened patch features $z \in \mathbb{R}^{HW \times C_{in}}$, as:

$$
\hat{z}_{(k,c)} = \sum_{k=1}^{HW} \sum_{c=1}^{C_{in}} w(k,c) \cdot z(k + \Delta_{DMLP}^{(k,c),c}),
$$

where $w \in \mathbb{R}^{C_{in} \times C_{out}}$ is the weight matrix of a fully-connected (FC) layer. As shown in Fig 2c, the decoder has a similar structure as a MLP-Mixer block [60], consisting of DPE, DMLP, and MLP modules. The residual connections are kept. Formally, the four-stage decoder is denoted as:

$$
\begin{align*}
\hat{z}_l &= DPE(C_l, C_{emb}) (\hat{z}_l), \forall l \in \{1, 2, 3, 4\} \\
\hat{z}_l &= DMLP(C_{emb}, C_{emb}) (\hat{z}_l) + \hat{z}_l, \forall l \\
\hat{z}_l &= MLP(C_{emb}, C_{emb}) (\hat{z}_l) + \hat{z}_l, \forall l \\
\hat{z}_l &= Up(H/4, W/4) (\hat{z}_l), \forall l \\
p &= LN(C_{emb}, C_K)(\sum_{l=1}^{4} \hat{z}_l),
\end{align*}
$$

where $Up(\cdot)$ and $LN(\cdot)$ refer to the Upsample- and Layer-Norm operations, and $p$ is the prediction of $K$ classes.

### 3.4. Mutual Prototypical Adaptation

Due to the lack of large-scale training data in panoramas, we look into Pin2PAN domain adaptation from a perspective of semantic prototypes [91]. We propose the **Mutual Prototypical Adaptation (MPA)** method to enable distilling knowledge via prototypes which we cultivate through source ground truth labels and target pseudo labels. Pseudo-labels depend on the few remaining mutual properties from pinhole and panoramic images, e.g., scene distribution at the frontal viewing angle [9, 75]. While the related PCS [84] performs inter- and intra-domain instance-prototype learning, our mutual prototypes are learned from source- and target feature embeddings $f^s$ and $f^t$, projected to a shared latent space, and stored in a dynamic bank, as shown in Fig. 4. The key differences to PCS lie in that (1) the mutual prototypes are built by joining embeddings from both domains, and (2) our method leverages multi-scale pyramidal features using different input scales in computing the embeddings which yields more robust prototypes.
Given the source (pinhole) dataset with annotated images $\mathcal{D}^s = \{(x^s_i, y^s_i)|x^s_i \in \mathbb{R}^{H \times W \times 3}, y^s_i \in \{0, 1\}\}$ and the target (panoramic) dataset $\mathcal{D}^t = \{(x^t_i)|x^t_i \in \mathbb{R}^{H \times W \times 3}\}$ without annotations, the goal of domain adaptation is to learn semantics from the source domain and transfer it to the target domain with $K$ shared classes. The network is trained in $\mathcal{D}^s$ based on the segmentation loss:

$$L^s_{SEG} = - \sum_{i,j,k=1}^{H,W,K} y^s_{i,j,k} \log(p^s_{i,j,k}),$$

where $p^s_{i,j,k}$ indicates the probability of pixel $x^s_{i,j}$ predicted as $k$-th class on the source domain. To generalize the source pre-trained model to the target data, a typical Self-Supervised Learning (SSL) scheme optimizes the model based on the pseudo labels $\hat{y}^s_{i,j,k}$ of pixels $x^s_{i,j}$ in the target domain:

$$L^t_{SSL} = - \sum_{i,j,k=1}^{H,W,K} \hat{y}^s_{i,j,k} \log(p^s_{i,j,k}),$$

where the pseudo label is given by the most probable class in the model predictions: $\hat{y}^s_{i,j,k} = \mathbb{1}_{k=\arg\max p^s_{i,j,k}}$. However, training with hard pseudo-labels leaves the model sensitive and fragile against errors in its own prediction and has only a limited positive effect on performance. Therefore, we advocate prototype-based alignment in the feature space, which brings two benefits: (1) it softens the hard pseudo-labels by using them in feature space instead of as direct targets and (2) it performs complementary alignment of semantic similarities in feature space.

Specifically, given a set with all $n_s$ source- and $n_t$ target feature maps $F = \{f^1_t, \ldots, f^n_t\} \bigcup \{f^1_s, \ldots, f^n_s\}$, with feature maps $f$ fused from four-stage multi-scale features $f = \sum_{i=1}^4 f_i$. Each feature map is associated either with its respective source ground-truth label or a target pseudo-label. To compute the mutual prototype memory $\mathcal{M} = \{P_1, \ldots, P_K\}$ with prototypes $P_k$ we take the mean of all feature vectors (pixel-embeddings) from all feature maps in $F$ that share the class label $k$. We initialize $\mathcal{M}$ by computing the class-wise mean embeddings through the whole dataset and while training we update the prototype $P_k$ at timestep $t$ online by $P_{k}^{t+1} \leftarrow m P_{k}^{t-1} + (1 - m) P_{k}^{t}$ with a momentum $m=0.999$, where $P_{k}^{t}$ are the mean pixel-embedding among embeddings that share the class-label $k$ in the current mini-batch. An overview of this procedure is displayed in Fig. 4. The mutual prototypical adaptation loss is inspired by the knowledge distillation loss [6], which drives the feature embedding $\hat{f}$ to be aligned with the prototypical feature map $f$ which is set up, by stacking the prototypes $P_k \in \mathcal{M}$ according to the pixel-wise class distribution in either the source label or the pseudo-label. The resulting target $\hat{f}$ has the same shape as $f$. For brevity, only the source domain is displayed in Eq. (6), which is similar to the target domain.

$$L^s_{MPA} = - \lambda T^2 \text{KL}(\phi(\hat{f}^s/T)||\phi(f^s/T)) - (1 - \lambda) \text{CE}(y^s, \phi(f^s)),$$

where $\text{KL}(-)$, $\text{CE}(-)$, and $\phi(-)$ are Kullback–Leibler divergence, Cross-Entropy, and Softmax function, respectively. The temperature $T$ and hyper-parameter $\lambda$ are 20 and 0.9 in our experiments.

The final loss is combined with a weight of $\alpha=0.001$ as:

$$L = L^s_{SEG} + L^t_{SSL} + \alpha (L^s_{MPA} + L^t_{MPA}).$$

4. Experiments

4.1. Datasets and Settings

**Indoor pin(hole) dataset.** Stanford2D3D [1] (SPin for short) has 70,496 pinhole images. The dataset is collected in indoor areas and annotated with 13 categories. Results are averaged over 3 official folds, unless otherwise stated.

**Indoor pan(oramic) dataset.** Stanford2D3D [1] (SPan for short) has 1,413 panoramic images. The images are annotated with the same 13 categories as its pinhole dataset.

**Outdoor pin(hole) dataset.** Cityscapes [11] (CS for short) dataset comprises 2,979 and 500 images for training and validation. Images are annotated with 19 categories.

**Outdoor pan(oramic) dataset.** DensePASS [45] (DP for short) collected from cities around the world has 2,000 images for transfer optimization and 100 labeled images for testing, annotated with the same 19 classes as Cityscapes.

**Implementation settings.** We train Trans4PASS models with 4 1080Ti GPUs with an initial learning rate of $5e-5$, scheduled by the poly strategy with power 0.9 over 200 epochs. AdamW [31] is the optimizer with epsilon 1e−8, weight decay $1e-4$ and batch size is 4 on each GPU. The image augmentations include random resize with ratio 0.5–2.0, random horizontal flipping, and random cropping to $512 \times 512$. For outdoor datasets, the resolution is $1080 \times 1080$ and batch size is 1. When adapting the models from Pin2Pan, the resolution of indoor pinhole and panoramic images are $1080 \times 1080$ and $1024 \times 512$ for training, while the outdoor images are set to $1024 \times 512$ and $2048 \times 400$. The image size of indoor and outdoor validation
To quantify the Domain gap in outdoor scenarios. images from Stanford2D3D are captured under the same domain gaps in indoor scenarios. As pinhole and panoramic Domain gap in indoor scenarios. are important for wide-FoV panoramic segmentation. indicates that distortion-aware features and long-range cues are maintained in both low and high levels of Transformers as opposed to the context learned in higher-levels of CNNs, are important for wide-FoV panoramic segmentation.

**Domain gap in outdoor scenarios.** To quantify the PIN2PAN domain gap in outdoor scenarios, we evaluate over 15 off-the-shelf segmentation models trained on Cityscapes. Table 1 summarizes the results tested on Cityscapes and DensePASS validation sets. Although previous transformers [68, 94] reduce the mIoU gap from CNN-based counterparts to ~40%, the PIN2PAN gap remains large. The proposed Trans4PASS architecture has a high performance on pinhole image segmentation and also outperforms other methods on panoramic segmentation with 44.8% mIoU without any adaptation strategy. It indicates that distortion-aware features and long-range cues maintained in both low and high levels of Transformers as opposed to the context learned in higher-levels of CNNs, are important for wide-FoV panoramic segmentation.

**Domain gap in indoor scenarios.** Table 2 shows PIN2PAN domain gaps in indoor scenarios. As pinhole and panoramic images from Stanford2D3D are captured under the same setting, the PIN2PAN gap is smaller compared to the outdoor scenario. Still, in light of other CNN- and transformer-based methods, the small Trans4PASS version achieves 50.20% and 48.34% mIoU in pinhole- and panoramic image segmentation, yielding the smallest performance drop.

### 4.3. Trans4PASS Structural Analysis

**Effect of DPE.** We compare DPE against DePatch from DPT [7]. While the object-aware offsets and scales in DPT make patches shift around the object, our DPE is flexible to split image patches and is decoupled from object proposals. As shown in the first group of Table 3, compared with DPT, our DPE-based Trans4PASS adds +3.01% and +9.39% mIoU on Cityscapes and DensePASS, respectively.

**Effect of DMLP.** To ablate the effect of different MLP-like modules embedded in the decoder of Trans4PASS, we substitute DMLP by CycleMLP [5] and ASMLP [36] modules. DMLP is lighter than ASMLP with fewer GFLOPs, parameters and it is more adaptive as opposed to the fixed offsets in CycleMLP. The first group of Table 3 shows that DMLP outperforms both modules with 3% to 5% in mIoU.

**Effect of encoders and decoders.** With the same encoder and decoder paired with DPT, our DPE-based Trans4PASS adds +3.01% and +9.39% mIoU on Cityscapes and DensePASS, respectively.
### 4.4. PIN2PAN Adaptation

**Ablations in outdoor scenarios.** To verify the generalization ability of applying TransPASS in adaptation methods, Fanet and DANet used in P2PDA [45] are replaced by TransPASS-T/-S, as visible in Table 4b. TransPASS brings >10% performance gains due to the captured long-range contexts and distortion-aware features. Without the advantage of a superior network architecture, MPA achieves 51.93% and 54.77% with TransPASS-T and -S models, surpassing 51.05% and 52.91% of P2PDA. The second and third ablation groups of Table 4b show how TransPASS-T and -S match up against each other. Individually, MPA is paired with the SSL-based method. When combining both, MPA and SSL, TransPASS-S obtains new state-of-the-art performance on DensePASS, reaching 55.25% in mIoU and 56.38% with multi-scale evaluation. This verifies that MPA works collaboratively with pseudo labels and provides a complementary feature alignment incentive.

**Omnidirectional segmentation.** To showcase the effectiveness of MPA on omnidirectional segmentation, the panoramic image is divided into 8 directions and evaluated individually. The polar diagram in Fig. 3 demonstrates that MPA brings uniform improvement to omnidirectional seg-

![Figure 5: Comparison of omnidirectional segmentation before and after mutual prototypical adaptation.](image)

Apart from benefiting the stuff classes (road, sidewalk, and terrain), MPA improves the segmentation of object classes, such as person and truck. Due to the panorama boundary at 180°, InUs of motorcycle and bicycle are impacted, still consistent and large accuracy boosts with MPA in all directions for different classes are observed.

**Comparison with outdoor state-of-the-art methods.** In Table 4a, we compare our solution with recent panoramic segmentation [72, 75] and domain adaptation [44, 67, 84, 88, 97], and multi-supervision methods [30, 50, 90]. * denotes performing multi-scale (MS) evaluation.

### Table 4. Comparisons and ablation studies of PIN2PAN domain adaptation in indoor and outdoor scenarios.

| Network | Method | mIoU(%) |
|---------|--------|---------|
| FA-Net | P2PDA  | 33.67   |
| DANet  | P2PDA  | 43.99   |
| TransPASS-T |  | 51.05   |
| TransPASS-S |  | 52.91   |
| TransPASS-T | Warm-up | 45.89   |
| TransPASS-T | SSL    | 50.56   |
| TransPASS-S | SSL    | 51.86   |
| TransPASS-S | MPA    | 51.93   |
| TransPASS-T | MPA + SSL | 53.26   |
| TransPASS-S | MPA + SSL + MS | 54.72   |
| TransPASS-T | -       | 48.73   |
| TransPASS-S | Warm-up | 52.59   |
| TransPASS-S | SSL    | 54.67   |
| TransPASS-S | MPA    | 54.77   |
| TransPASS-S | MPA + SSL | 55.28   |
| TransPASS-S | MPA + SSL + MS | 56.38   |

| Network | Method | mIoU(%) |
|---------|--------|---------|
| DANet  | -      | 40.28   |
| TransPASS-T | P2PDA  | 42.62   |
| PVT-Tiny | -      | 24.45   |
| PVT-Tiny | P2PDA  | 39.66   |
| PVT-Small | -   | 23.11   |
| PVT-Small | P2PDA | 43.10   |
| Trans4PASS-T | -  | 46.08   |
| Trans4PASS-S | MPA    | 47.48   |
| Trans4PASS-S | -     | 48.34   |
| Trans4PASS-S | MPA    | 52.15   |
| DANet  | Supervised | 44.15   |
| Trans4PASS-S | Supervised | 53.31   |

| Method | Input | mIoU(%) |
|--------|-------|---------|
| SuDD  | RGB   | 32.6    |
| CubeMap | RGB  | 33.8    |
| DustMap | RGB  | 34.6    |
| LUNA [53] | RGB-D | 35.9    |
| GauzeNet [10] | RGB-D | 39.4    |
| UGSCNN [28] | RGB-D | 38.3    |
| HexRUNet [86] | RGB-D | 43.3    |
| Tangent [16] (ResNet-101) | RGB  | 45.6    |
| HoloeNet [58] (ResNet-101) | RGB  | 52.0    |
| Trans4PASS (Small) | RGB  | 52.1    |
| Trans4PASS (Small+MS) | RGB  | 53.0    |
| Trans4PASS (Source only) | RGB  | 48.1    |
| Trans4PASS (MPA) | RGB  | 50.8    |
| Trans4PASS (MPA+MS) | RGB  | 51.2    |

(a) Per-class results on DensePASS. Comparison with state-of-the-art panoramic segmentation [72, 75], domain adaptation [44, 67, 84, 88, 97], and multi-supervision methods [30, 50, 90]. * denotes performing multi-scale (MS) evaluation.

(b) Adaptation results on DensePASS.

(c) Adaptation results on SPan @ fold-1.

(d) Comparison on SPan averaged by 3 folds.
88, 97] methods. Following [88], we also involve multi-supervision methods [30, 50, 90] which require much more data, to broaden the comparison. MPA-Trans4PASS arrives at the highest mIoU of 56.38%, outperforming the previous best P2PDA-SSL on DensePASS by 14.39% and the prototypical method [84] adapted by Trans4PASS. Trans4PASS obtains top scores on 10 of 19 classes. Notably, our solution shows improvements on challenging categories, e.g., truck, train, motorcycle, and bicycle.

Adaptation results in indoor scenarios. The experiments in Table 4c are conducted according to the fold-1 data splitting [1] on the Stanford-Panoramic dataset. Our MPA surpasses the previous state-of-the-art P2PDA with DANet and it is even better than the one adapted by a PVT-Small backbone. Overall, our Trans4PASS-S with MPA achieves the highest mIoU (52.15%), even reaching the level of the fully-supervised Trans4PASS-S (53.31%) which does have access to panoramic image annotations.

Comparison with indoor state-of-the-art methods. Before and after adaptation in Table 4d, our Trans4PASS-S model (~14M parameters) obtains a high mIoU score (51.2%), even comparable to existing fully-supervised and transfer-learning methods, which are based on ResNet-101 backbones (~44M parameters and 52.0% mIoU).

4.5. Qualitative analysis

Panoramic semantic segmentation visualizations. Fig. 6a and Fig. 6c demonstrate that Trans4PASS handles the distortion of panoramic images very well as compared to indoor [65] and outdoor [68] baseline models. Especially, the segmentation results for sidewalks and pedestrians from Trans4PASS have more accurate classifications and boundary distinctions, while the baseline model is confused by the distorted shape and space, due to the lacking capacity to learn long-range contexts and distortion-aware features. In the indoor case of Fig. 6c, the door and chair categories are barely detected by the baseline model, but our Trans4PASS can output precise segmentation masks on both objects.

DPE and DMLP visualizations. Fig. 6b and Fig. 6d visualize effects of Deformable PE from four stages of Trans4PASS. The red dots denote the centers of a selected patch (size of $s \times s$) sequence. Given learned offsets from DPE, $s^2$ yellow sampling dots are shifted to semantic-relevant areas in a flexible way, where each pixel is adaptive to distorted objects and space, like the deformed building and sidewalk (see Stage-4 DPE in Fig. 6b). Besides, to verify the effect of Deformable MLP, two feature map pairs from the 75th channel before and after DMLP are displayed in Fig. 6e and 6e. The feature maps (indoors/outdoors) after DMLP present semantically recognizable responses, e.g., on regions of distorted sidewalks or doors, as compared to those before the DMLP module.

5. Conclusion

To revitalize 360° scene understanding, we introduce a universal framework with a Transformer for Panoramic Semantic Segmentation (Trans4PASS) model and a Mutual Prototypical Adaptation (MPA) method for transferring semantic information from the label-rich pinhole domain to the label-scarce panoramic domain. The Deformable Patch Embedding (DPE) and the Deformable MLP (DMLP) module endow Trans4PASS with distortion awareness. The framework elevates state-of-the-art performances on the competitive Stanford2D3D and DensePASS benchmarks.

Limitations. We note that the accuracy of some classes are still impacted by the partition boundary of panoramas at 180°. Transferring models between pinhole-, fisheye-, and panoramic domains, fusing modalities, and solving various tasks of 360° imagery are opportunities for further research.
References

[1] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105, 2017. 2, 5, 8, 12, 13, 14

[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 CVPR, 2019. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018. 6

[4] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In ICCV, 2019. 2

[5] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. CycleMLP: A MLP-like architecture for dense prediction. In ICLR, 2022. 6

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In NeurIPS, 2020. 5

[7] Zhiyang Chen, Yousong Zhu, Chaoqyang Zhao, Guosheng Hu, Wei Zeng, Jingqiao Wang, and Ming Tang. DPT: Deformable patch-based transformer for visual recognition. In MM, 2021. 4, 6

[8] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In ICCV, 2021. 2

[9] Sungha Choi, Joanne T. Kim, and Jaegul Choo. Cars can’t fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In CVPR, 2020. 4

[10] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In ICML, 2019. 1, 7, 14

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Marku Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In IVC, 2016. 5

[12] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In ICCV, 2017. 3, 4, 12

[13] Grégoire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P. Breckon. Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360° panoramic imagery. In ECCV, 2018. 1

[14] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using round view cameras. T-ITS, 2020. 2, 12

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021. 2, 3

[16] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In CVPR, 2020. 7, 14

[17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In CVPR, 2019. 1, 2, 6, 12

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NeurIPS, 2014. 2

[19] Qi Qi Gu, Qianyu Zhou, Minhao Xu, Zhengyang Feng, Guangliang Cheng, Xuequan Lu, Jiaping Shi, and Lizhuang Ma. PIT: Position-invariant transform for cross-FoV domain adaptation. In ICCV, 2021. 2

[20] Julia Guerrero-Viiu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J. Guerrero. What’s in my room? Object recognition on indoor panoramic images. In ICRA, 2020. 1

[21] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. MetaCorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In CVPR, 2021. 2

[22] 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 ICML, 2018. 2

[23] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In CVPR, 2020. 2

[24] Ping Hu, Federico Perazzi, Fabian Caba Heilbron, Oliver Wang, Zhe Lin, Kate Saenko, and Stan Sclaroff. Real-time semantic segmentation with fast attention. RA-L, 2021. 6, 12

[25] Jiaxing Huang, Shijian Lu, Dayan Guan, and Xiaobing Zhang. Contextual-relation consistent domain adaptation for semantic segmentation. In ECCV, 2020. 2

[26] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-cross attention for semantic segmentation. In ICCV, 2019. 2

[27] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. In IV, 2021. 2

[28] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, and Matthias Nießner. Spherical CNNs on unstructured grids. In ICLR, 2019. 2, 7, 14

[29] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In ICCV, 2021. 2

[30] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C. V. Jawahar. Universal semi-supervised semantic segmentation. In ICCV, 2019. 7, 8

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 5

[32] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In CVPR, 2019. 6
[33] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. SpherePHD: Applying CNNs on a spherical PolyHeDron representation of 360° images. In CVPR, 2019. 2

[34] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In ECCV, 2020. 2

[35] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In CVPR, 2019. 2

[36] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. AS-MLP: An axial shifted MLP architecture for vision. arXiv preprint arXiv:2107.08591, 2021. 2, 6

[37] Qing Lian, Lixin Duan, Fengmao Lv, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In ICCV, 2019. 2

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, 2017. 4

[39] Hanxiao Liu, Zhiyang Dai, David R. So, and Quoc V. Le. Pay attention to MLPs. In NeurIPS, 2021. 2

[40] Mengyi Liu, Shuhui Wang, Yulan Guo, Yuan He, and Hui Xue. Pano-SM Learner: Self-Supervised multi-task learning of depth and semantics in panoramic videos. SPL, 2021. 2

[41] Yahao Liu, Jinhong Deng, Xincheng Gao, Wen Li, and Lixin Duan. BAPA-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In ICCV, 2021. 2

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Xixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021. 2

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015. 2

[44] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversarial training for semantic consistent domain adaptation. In CVPR, 2019. 2, 7, 8

[45] 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 ITSC, 2021. 1, 2, 3, 5, 7, 12, 13

[46] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In CVPR, 2021. 2

[47] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. JIVP, 2021. 2

[48] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images. In CVPR, 2019. 6, 7

[49] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In CVPR, 2020. 2

[50] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In CVPR, 2019. 7, 8

[51] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-SCNN: Fast semantic segmentation network. In BMVC, 2019. 6

[52] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. T-ITS, 2018. 6, 7

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In MICCAI, 2015. 7, 14

[54] Mehran Shakerinava and Siamak Ravanbakhsh. Equivariant networks for pixelized spheres. In ICML, 2021. 1, 2

[55] Prabhu Teja Sivaprasad and François Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In CVPR, 2021. 2

[56] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In ICCV, 2021. 2

[57] Yi Yang. Content-consistent matching for domain adaptive semantic segmentation through self-supervision. In CVPR, 2019. 2

[58] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schuler, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In CVPR, 2018. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 2

[60] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In ICML, 2021. 2

[61] Xiulong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, Lukasz Kaiser, and Illia Polosukhin. MLP: An axial shifted MLP architecture for vision. In NeurIPS, 2021. 2
A. Quantitative analysis

A.1. Analysis of hyper-parameters

As the spatial correspondence problem indicated in [12], if the deformable convolution is applied to the lower or middle layers, the spatial structures are susceptible to fluctuation [14]. To overcome this problem, we propose the regional restriction of learned offsets to stabilize the training of our early-stage and four-stage Deformable Patch Embedding (DPE) module. Table 5 shows that \( r=4 \) has a better result. Thus, the constraint \( r \) applied in the offset prediction module is set as 4 in our experiments.

To investigate the effect of various hyper-parameters in the proposed Trans4PASS framework, we analyze the weight \( \alpha \) and the temperature \( T \) as shown in Fig. 7a and Fig. 7b. The weight \( \alpha \) is used to combine the Mutual Prototypical Adaptation (MPA) loss and the source- and target segmentation losses. As \( \alpha \) decreases from 0.1 to 0, we set the temperature \( T=35 \) in the MPA loss and evaluate the mIoU(%) results on the target (DensePASS [45]) dataset. If \( \alpha=0 \), the final loss is equivalent to that of the SSL-based method, i.e., the MPA loss is excluded. When \( \alpha=0.001 \) for combining both, MPA and SSL, Trans4PASS obtains a better performance.

Apart from the combination weight \( \alpha \), we further investigate the effect of the temperature \( T \), which is used in the MPA loss. As shown in Fig. 7b, the performance is not sensitive to the distillation temperature, which illustrates the robustness of our MPA method. Nevertheless, we found that MPA performs better when the temperature is lower, so \( T=20 \) is set as the default setting in our experiments.

A.2. Computational complexity

We reported the complexity of Deformable Patch Embedding (DPE) and Deformable MLP (DMLP) and compared with other methods on DensePASS in Table 6. The results indicate that our methods have significant improvement with the same order of complexity.

A.3. Detailed results in outdoor scenarios

Table 7 shows the per-class IoU results on DensePASS dataset. The first group of experiments is conducted to compare the performance of different backbones in P2PDA [88] method. Additionally, the adaptation process of the original FANet [24] and DANet [17] are shown in more detail, i.e., the performance of the source-only model and that without using the SSL-based method are included. The experiments in the second and third groups are based on Trans4PASS-T and -S model, respectively. As shown in the third group, Trans4PASS-S obtains new state-off-the-art performance in mean IoU (56.38%). In addition, it achieves top scores on 7 out of 19 classes in per-class IoU, including pole, traffic light, person, car, truck, motorcycle, and bicycle.

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{None} & r=1 & r=2 & r=4 & r=8 \\
\hline
\text{mIoU(\%)} & 45.74 & 45.14 & 45.59 & 45.89 & 45.57 \\
\hline
\end{array}
\]

Table 5. Effect of regional restriction (\( r \)) on DensePASS.

Figure 7: Analysis of hyper-parameters. The performance (mIoU) is evaluated in the outdoor target dataset (DensePASS).

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{GFLOPs} & \text{\# Params(M)} & \text{DPE} & \text{CycleMLP(\%)} & \text{ASMLP(\%)} & \text{DMLP} \\
\hline
\text{EPE[79]} & 0.16 & 0.01 & 7.65 & 45.14 & 45.89 & 45.57 \\
\text{DPT[8]} & 0.36 & 0.02 & 2.90 & 38.50 & 40.16 & 42.05 \\
\text{ASMLP[43]} & 1.25 & 0.45 & 1.04 & 42.05 & 45.89 & 42.05 \\
\hline
\end{array}
\]

Table 6. Computational complexity of DPE and DMLP. GFLOPs are calculated @512×512.

A.4. Detailed results in indoor scenarios

Apart from the detailed results in outdoor scenarios, per-class results on the outdoor Stanford2D3D-Panoramic dataset [1] are shown in Table 8. The experiments are conducted on the fold-1 dataset setting of Stanford2D3D [1]. Our proposed framework with the Trans4PASS-S backbone and the MPA method obtains the best performance in the domain adaptation setting, reaching 52.15% in mean IoU. It also achieves best IoU scores on 7 out of 13 classes in the indoor scenario, especially on the ceiling, column, and door categories. In the supervised learning setting, Trans4PASS-S surpasses the CNN-based DANet by a large margin, achieving a score of 53.31% in mean IoU. Besides, its performance in per-class IoU is better than DANet in almost all categories, which lacks the capacity to learn long-range contexts and distortion-aware features in panoramas.

The comparison of segmentation performance with state-of-the-art methods on Stanford2D3D-Panoramic dataset is shown in Table 9. Since the results of these experiments are based on the average of all 3 data-splitting settings, we show the results of each individual split setting and its per-class IoU in detail (in gray). The small version of the Trans4PASS backbone is used in this experiment. Compared with the previous best fully-supervised method equipped with ResNet-101, Trans4PASS-S has much fewer parameters and is an order of magnitude smaller than ResNet-101. Still, our method obtains the new state-of-the-art performance on Stanford2D3D-Panoramic dataset, reaching 53.0% in mean IoU. Within all 13 classes, Trans4PASS obtains a total of 8 best per-class IoUs. In the
setting of unsupervised domain adaptation (UDA), our proposed method achieves +2.7\% in the average of three folds, and +3.1\% when using multi-scale evaluation. It obtains best per-class scores on 9 out of 13 categories.

B. Qualitative analysis

B.1. More visualizations in indoor scenarios

Similar to the visualization in outdoor scenarios, more qualitative comparisons between the baseline and the proposed Trans4PASS are displayed in Fig. 8, which are from the evaluation set of Stanford2D3D-Panoramic [1] in the fold-1 setting. In Fig. 8(a), Trans4PASS can produce higher quality segmentation results in those categories highlighted by the black dashed rectangles, such as column and bookcase categories, while the baseline model can hardly identify these severely deformed objects. In Fig. 8(b), the doors are incorrectly segmented as part of the wall by the baseline model, and the correct segmentation results can be generated by our Trans4PASS model.

B.2. More visualizations in outdoor scenarios

To fully demonstrate the effect of Trans4PASS in dealing with image distortions and object deformations, more qualitative comparisons between the baseline and the proposed Trans4PASS are displayed in Fig. 9, which are generated from the evaluation set of DensePASS dataset [45]. Specifically, Trans4PASS can better classify and segment deformed foreground objects with accurate boundaries, such as the segmentation results of cars and trucks highlighted by the blue dashed rectangles in Fig. 9(a), while the baseline model without deformable PE and deformable MLP modules is likely to be confused or fail in these categories. Apart from the foreground object, the ultra-wide arranged background is particularly distorted and challenging. Thanks to the two distortion-aware modules, our Trans4PASS yields high-quality segmentation results in these categories, e.g., terrain, sidewalk, and wall in Fig. 9(b).

C. Broader Impact.

This work promotes panoramic semantic segmentation of indoor and outdoor scenes, which benefits ultra-wide scene understanding. However, the proposed method has not been verified in practical applications such as those in intelligent vehicles and mobility assistive systems. As the experiments are conducted based on the referred datasets, there are still data biases in different test fields. If the learned model is directly applied to real scenarios, it may cause negative social impacts such as less reliable decision with less accurate segmentation, which should be considered in the downstream applications.
Table 7. Per-class results on DensePASS dataset. 'SSL' represents the self-supervised learning with pseudo-labels. '-' means no adaptation. 'MS' denotes multi-scale evaluation.

| Network  | Method          | beam | board | building | ceiling | chair | column | curtain | desk | door | table | traffic light | wall | car | motorcycle | bicycle |
|----------|-----------------|------|-------|----------|---------|-------|--------|---------|------|------|-------|---------------|------|-----|------------|---------|
| DANet    | P2PDA           | 30.8 | 44.7  | 40.6     | 15.2    | 50.1  | 11.5   | 68.3    | 29.8 | 29.4 | 47.7  | 42.1          | 50.1 | 41.5| 39.1       | 37.3    |
| Trans4PASS-T | -               | 36.0 | 46.9  | 40.7     | 15.5    | 50.4  | 10.9   | 68.5    | 29.8 | 28.6 | 47.1  | 42.3          | 49.2 | 41.5| 38.6       | 36.7    |
| Trans4PASS-T | Warm-up         | 37.3 | 47.3  | 40.9     | 15.5    | 50.7  | 10.9   | 68.6    | 29.8 | 28.4 | 47.4  | 42.3          | 49.4 | 41.7| 38.7       | 36.9    |
| Trans4PASS-T | MLA + SSL       | 37.3 | 47.3  | 40.9     | 15.5    | 50.7  | 10.9   | 68.6    | 29.8 | 28.4 | 47.4  | 42.3          | 49.4 | 41.7| 38.7       | 36.9    |
| Trans4PASS-T | MLA + SSL + MS  | 37.3 | 47.3  | 40.9     | 15.5    | 50.7  | 10.9   | 68.6    | 29.8 | 28.4 | 47.4  | 42.3          | 49.4 | 41.7| 38.7       | 36.9    |

Table 8. Per-class results on Stanford2D3D-Panoramic dataset according to the fold-1 data setting [1].

| Method      | Input | beam | board | building | ceiling | chair | column | curtain | desk | door | table | traffic light | wall | car | motorcycle | bicycle |
|-------------|-------|------|-------|----------|---------|-------|--------|---------|------|------|-------|---------------|------|-----|------------|---------|
| StdConv [59] | RGB   | 33.6 | 45.1  | 39.2     | 14.9    | 49.8  | 10.3   | 68.3    | 29.8 | 28.5 | 47.1  | 42.3          | 49.4 | 41.7| 38.7       | 36.9    |
| CubeNet [59] | RGB   | 33.8 | 45.3  | 39.1     | 14.6    | 49.5  | 10.1   | 68.4    | 29.7 | 28.4 | 47.2  | 42.2          | 49.3 | 41.5| 38.5       | 36.7    |
| U-Net [53]  | RGB-D | 35.9 | 47.2  | 40.9     | 15.2    | 50.3  | 11.0   | 69.5    | 30.0 | 28.8 | 47.6  | 42.5          | 49.5 | 41.8| 38.9       | 37.0    |
| SegNet [10] | RGB   | 31.9 | 43.2  | 38.8     | 14.2    | 48.3  | 9.9    | 67.7    | 28.5 | 27.7 | 46.4  | 41.9          | 48.8 | 41.1| 37.2       | 35.4    |
| UGSCNN [28] | RGB-D | 34.9 | 46.1  | 39.5     | 15.0    | 49.6  | 10.8   | 68.4    | 29.7 | 28.4 | 47.2  | 42.3          | 49.3 | 41.5| 38.5       | 36.7    |
| HexUENet [36] | RGB-D | 33.1 | 45.0  | 39.1     | 14.7    | 49.5  | 10.1   | 68.4    | 29.7 | 28.4 | 47.2  | 42.2          | 49.3 | 41.5| 38.5       | 36.7    |
| ResNet-101  | RGB   | 35.8 | 47.2  | 40.9     | 15.2    | 50.3  | 11.0   | 69.5    | 30.0 | 28.8 | 47.6  | 42.5          | 49.5 | 41.8| 38.9       | 37.0    |
| HolNet (ResNet-101) [58] | RGB | 52.0 | 46.1  | 47.7     | 21.7    | 48.7  | 10.9   | 68.5    | 29.9 | 28.8 | 47.6  | 42.5          | 49.5 | 41.5| 38.9       | 37.0    |

Table 9. Comparison on Stanford2D3D-Panoramic dataset. 'F-' is the result of the fold-1 (in gray) setting of Stanford2D3D [1]. 'Avg' is the averaged result of all 3 folds. 'MS' is multi-scale evaluation. 'UDA' is short for unsupervised domain adaptation.
Figure 9: Qualitative comparisons in outdoor scenarios.