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

Jiaming Zhang[1] Kailun Yang[1*] Chaoxiang Ma[2] Simon Reiß[1,3] Kunyu Peng[1] Rainer Stiefelhagen[1]
[1]CV:HCI Lab, Karlsruhe Institute of Technology [2]ByteDance Inc. [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]

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].

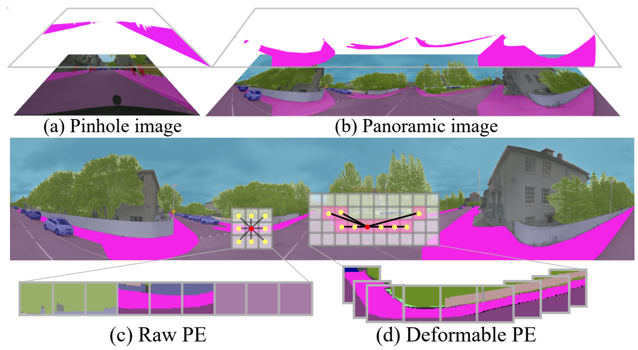


Figure 1: Semantic segmentation of (a) narrow-angle pinhole image and (b) 360° panoramic image. Compared to (c) standard Patch Embeddings, our (d) Deformable Patch Embedding partitions 360° images while considering distortions, *e.g.* in *sidewalks*.

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].

*Corresponding author (e-mail: kailun.yang@kit.edu).
[1]Code will be made publicly available at https://github.com/jamycheung/Trans4PASS.

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.

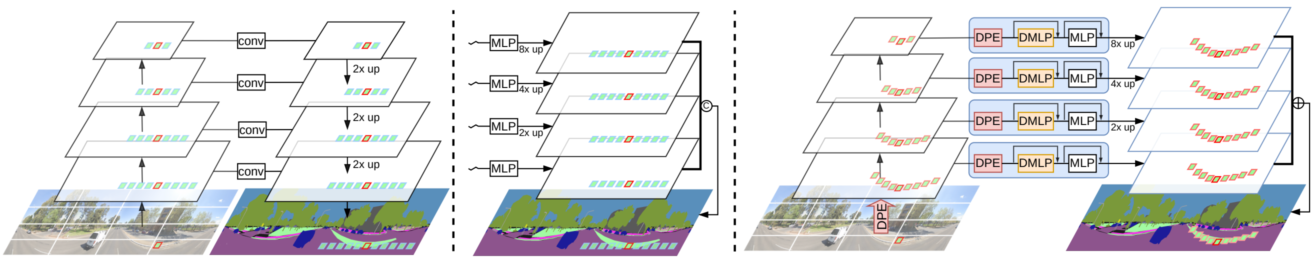
(3) Our framework for transferring models from PIN2PAN yields excellent results on two competitive benchmarks: On Stanford2D3D we circumvent using $1,400$ expensive panorama labels while achieving comparable results and on DensePASS we boost state-of-the-art performance by an absolute 14.39% in mIoU.

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



(a) Transformer with FPN-like decoder (b) Transformer with vanilla-MLP (c) Trans4PASS with DPE and DMLP

Figure 2: **Comparison of segmentation transformers.** Transformers (a) borrow a FPN-like decoder [94] from CNN counterparts or (b) adopt a vanilla-MLP decoder [68] for feature fusion, which lacks patch mixing. (c) *Trans4PASS* integrates Deformable Patch Embeddings (DPE) and the Deformable MLP (DMLP) module for capabilities to handle distortions (see warped *terrain*) and mix patches.

P2PDA [45] first tackles PIN2PAN transfer by learning attention correspondences. Aside from distortion-adaptive architecture design, we revisit PIN2PAN 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 feature embeddings 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 $\boldsymbol{f}_l{\in}\{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \boldsymbol{f}_4\}$ 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 $\frac{H}{4}{\times}\frac{W}{4}{\times}C_{emb}$, 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 $\boldsymbol{f}{\in}\mathbb{R}^{H\times W\times C_{in}}$, a standard PE module [15, 68] splits it into a flattened 2D patch sequence $\boldsymbol{z}{\in}\mathbb{R}^{(\frac{HW}{s^2})\times(s^2\cdot C_{in})}$, 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 $\boldsymbol{z}$ 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 $\boldsymbol{\Delta}_{(i,j)}{\in}\mathbb{N}^2$. In standard PE, these offsets are fixed and lie in $\boldsymbol{\Delta}_{(i,j)}{\in}[\lfloor-\frac{s}{2}\rfloor, \lfloor+\frac{s}{2}\rfloor]^2$. Take *e.g.* a 3×3 patch, offsets $\boldsymbol{\Delta}_{(i,j)}$ 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 $\boldsymbol{\Delta}^{DPE}{\in}\mathbb{N}^{H\times W\times 2}$ that can better cope with the spatial connections of objects, as present in distorted patches. DPE is learnable and predicts relative offsets based on the original input $\boldsymbol{f}$. The offset $\boldsymbol{\Delta}_{(i,j)}^{DPE}$ is calculated as depicted in Eq. (1).

$$\boldsymbol{\Delta}_{(i,j)}^{DPE}=\begin{bmatrix}\min(\max(-\frac{H}{r},g(\boldsymbol{f})_{(i,j)}),\frac{H}{r})\\ \min(\max(-\frac{W}{r},g(\boldsymbol{f})_{(i,j)}),\frac{W}{r})\end{bmatrix}, \tag{1}$$

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 $\boldsymbol{f}\in\mathbb{R}^{H\times W\times C_{in}}$, the spatial offset $\boldsymbol{\Delta}_{(i,j,c)}^{DMLP}$ is predicted channel-wise as in Eq. (1) and is

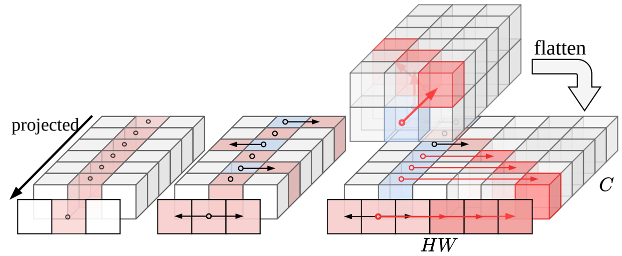


Figure 3: **Comparison of MLP blocks.** The spatial offsets of DMLP are learned adaptively from the input feature map.

then flattened as $\boldsymbol{\Delta}_{(k,c)}^{DMLP}$, where $k\in HW$ and $c\in C_{in}$, for mixing the flattened patch features $\boldsymbol{z}\in\mathbb{R}^{HW\times C_{in}}$, as:

$$\hat{\boldsymbol{z}}_{(k,c)}=\sum_{k=1}^{HW}\sum_{c=1}^{C_{in}}w_{(k,c)}^{T}\cdot\boldsymbol{z}_{(k+\boldsymbol{\Delta}_{(k,c)}^{DMLP},c)}, \tag{2}$$

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:

$$\begin{aligned}\hat{\boldsymbol{z}}_l&=\mathbf{DPE}(C_l,C_{emb})(\boldsymbol{z}_l),\forall_l\in\{1,2,3,4\}\\ \hat{\boldsymbol{z}}_l&=\mathbf{DMLP}(C_{emb},C_{emb})(\hat{\boldsymbol{z}}_l)+\hat{\boldsymbol{z}}_l,\forall_l\\ \hat{\boldsymbol{z}}_l&=\mathbf{MLP}(C_{emb},C_{emb})(\hat{\boldsymbol{z}}_l)+\hat{\boldsymbol{z}}_l,\forall_l\\ \hat{\boldsymbol{z}}_l&=\mathbf{Up}(H/4,W/4)(\hat{\boldsymbol{z}}_l),\forall_l\\ p&=\mathbf{LN}(C_{emb},C_K)(\sum_{l=1}\hat{\boldsymbol{z}}_l),\end{aligned} \tag{3}$$

where $\mathbf{Up}(\cdot)$ and $\mathbf{LN}(\cdot)$ refer to the Upsample- and LayerNorm 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 $\mathcal{D}^s{=}\{(x^s,y^s)|x^s{\in}\mathbb{R}^{H\times W\times 3}, y^s{\in}\{0,1\}^{H\times W\times K}\}$ and the target (panoramic) dataset $\mathcal{D}^t{=}\{(x^t)|x^t{\in}\mathbb{R}^{H\times W\times 3}\}$ 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:

$$\mathcal{L}^s_{SEG} = -\sum_{i,j,k=1}^{H,W,K} y^s_{(i,j,k)}\log(p^s_{(i,j,k)}), \tag{4}$$

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 $\hat{y}^t_{(i,j,k)}$ of pixels $x^t_{(i,j)}$ in the target domain:

$$\mathcal{L}^t_{SSL} = -\sum_{i,j,k=1}^{H,W,K} \hat{y}^t_{(i,j,k)}\log(p^t_{(i,j,k)}), \tag{5}$$

where the pseudo label is given by the most probable class in the model predictions: $\hat{y}^t_{(i,j,k)} = \mathbb{1}_{k\doteq\arg\max p^t_{(i,j,:)}}$. 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 $\boldsymbol{F}{=}\{\boldsymbol{f}^s_1,\dots,\boldsymbol{f}^s_{n_s}\}\bigcup\{\boldsymbol{f}^t_1,\dots,\boldsymbol{f}^t_{n_t}\}$, with feature maps $\boldsymbol{f}$ fused from four-stage multi-scale features $\boldsymbol{f}{=}\sum_{l=1}^{4} f_l$. Each feature map is associated either with its respective source ground-truth label or a target pseudo-label. To compute the mutual prototype memory $\boldsymbol{\mathcal{M}}{=}\{P_1,...,P_K\}$ with prototypes P_k we take the mean of all feature vectors (pixel-embeddings) from all feature maps in $\boldsymbol{F}$ that share the class label k. We initialize $\boldsymbol{\mathcal{M}}$ 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 $P_k^{t+1}{\leftarrow}mP_k^{t-1}{+}(1{-}m)P_k^t$ with a momentum $m{=}\ 0.999$, where P_k^t is 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 $\boldsymbol{f}$ to be aligned with the prototypical feature map $\hat{\boldsymbol{f}}$ which is set up, by stacking the prototypes $P_k{\in}\boldsymbol{\mathcal{M}}$ according to the pixel-wise class distribution in either the source label or the pseudo-label. The resulting target $\hat{\boldsymbol{f}}$ has the same shape as $\boldsymbol{f}$. For brevity, only the source domain is displayed in Eq. (6), which is similar to the target domain.

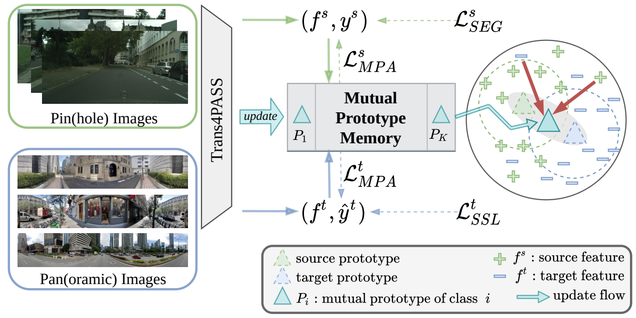


Figure 4: **Diagram of mutual prototypical adaptation.**

$$\begin{aligned}\mathcal{L}^s_{MPA} = &-\lambda\mathcal{T}^2\mathbf{KL}(\phi(\hat{\boldsymbol{f}}^s/\mathcal{T})||\phi(\boldsymbol{f}^s/\mathcal{T}))\\ &-(1-\lambda)\mathbf{CE}(y^s,\phi(\boldsymbol{f}^s)),\end{aligned} \tag{6}$$

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

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

$$\mathcal{L}{=}\mathcal{L}^s_{SEG}{+}\mathcal{L}^t_{SSL}{+}\alpha(\mathcal{L}^s_{MPA}{+}\mathcal{L}^t_{MPA}). \tag{7}$$

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

Network	Backbone	CS	DP	mIoU Gaps
SwiftNet [48]	ResNet-18	75.4	25.7	-49.7
Fast-SCNN [51]	Fast-SCNN	69.1	24.6	-44.5
ERFNet [52]	ERFNet	72.1	16.7	-55.4
FANet [24]	ResNet-34	71.3	26.9	-44.4
PSPNet [93]	ResNet-50	78.6	29.5	-49.1
OCRNet [81]	HRNetV2p-W18	78.6	30.8	-47.8
DeepLabV3+ [3]	ResNet-101	80.9	32.5	-48.4
DANet [17]	ResNet-101	80.4	28.5	-51.9
DNL [78]	ResNet-101	80.4	32.1	-48.3
Semantic-FPN [32]	ResNet-101	75.8	28.8	-47.0
ResNeSt [87]	ResNeSt-101	79.6	28.8	-50.8
OCRNet [81]	HRNetV2p-W48	80.7	32.8	-47.9
SETR-Naive [94]	Transformer-L	77.9	36.1	-41.8
SETR-MLA [94]	Transformer-L	77.2	35.6	-41.6
SETR-PUP [94]	Transformer-L	79.3	35.7	-43.6
SegFormer-B1 [68]	SegFormer-B1	78.5	38.5	-40.0
SegFormer-B2 [68]	SegFormer-B2	81.0	42.4	-38.6
Trans4PASS-T	Trans4PASS-T	79.1	41.5	-37.6
Trans4PASS-S	Trans4PASS-S	81.1	**44.8**	-36.3

Table 1. **Performance gaps** of CNN- and transformer-based models from Cityscapes (**CS**) @ 1024×512 to DensePASS (**DP**).

Network	Backbone	SPin	SPan	mIoU Gaps
Fast-SCNN [51]	Fast-SCNN	41.71	26.86	-14.85
SwiftNet [48]	ResNet-18	42.28	34.95	-7.87
DANet [17]	ResNet-50	43.33	37.76	-5.57
DANet [17]	ResNet-101	40.09	31.81	-8.28
Trans4Trans-T [89]	PVT-T	41.28	24.45	-16.83
Trans4Trans-S [89]	PVT-S	44.47	23.11	-21.36
Trans4PASS-T	Trans4PASS-T	49.05	46.08	-2.97
Trans4PASS-S	Trans4PASS-S	50.20	**48.34**	-1.86

Table 2. **Performance gaps** from Stanford2D3D-Pinhole (**SPin**) to Stanford2D3D-Panoramic (**SPan**) dataset on fold-1.

sets are 2048×1024 and 2048×400, respectively. Adaptation models are trained within $10K$ iterations on one GPU.

4.2. PIN2PAN Gaps

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.[1] Table 1 summarizes the results tested on Cityscapes and DensePASS validation sets. Although previous transformers [68, 94] reduce the mIoU gap from ∼50% of 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.

[1] MMSegmentation: https://github.com/open-mmlab/mmsegmentation.

Network	Encoder	Decoder	GFLOPs	#P	CS	DP
(1) Compare PEs and MLPs:						
Trans4PASS	MiT-B1*	DMLP	13.11	13.10	69.48	36.50
Trans4PASS	MiT-B1†	CycleMLP [5]	9.83	13.60	73.49	40.16
Trans4PASS	MiT-B1†	ASMLP [36]	13.40	14.19	73.65	42.05
Trans4PASS	MiT-B1†	DMLP	12.02	13.93	72.49	45.89 (+9.39)
(2) Compare encoders and decoders:						
PVT [65]	PVT-T	FPN	11.17	12.76	71.46	31.20
PVT [65]	PVT-T	Vanilla MLP	14.56	12.84	70.60	32.85
PVT [65]	PVT-T	DMLP	13.11	13.10	71.75	35.18 (+3.98)
Trans4PASS	PVT-T†	DMLP	13.18	13.10	69.62	36.50 (+5.30)
SegFormer [68]	MiT-B1	Vanilla MLP	13.27	13.66	74.93	39.02
SegFormer [68]	MiT-B1	FPN	9.88	13.58	73.96	41.14
SegFormer [68]	MiT-B1	DMLP	11.82	13.92	73.10	45.14 (+6.12)
Trans4PASS	MiT-B1†	DMLP	12.02	13.93	72.49	**45.89** (+6.87)

Table 3. **Trans4PASS structural analysis.** * and † denote DPT [7] and our DPE. "**#P**" is short for #Parameters in millions. Models are trained on Cityscapes (**CS**) @ 512×512 and tested on DensePASS (**DP**) @ 2048×400.

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 as PVT, a DMLP-based decoder brings a +3.98% improvement compared to the FPN- and MLP-based decoders, as shown in the second group of Table 3. When our DPE is applied in the early stage of the PVT encoder, further improvements of +5.30% can be made. Similar improvement results (+6.12% and +6.87%) are evident in experiments with a SegFormer encoder. Overall, these results show that DPE and DMLP can be integrated into diverse backbones, significantly improving distortion-adaptability for panoramic scene segmentation.

Method	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
ERFNet [52]	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) [72]	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
ECANet (Omni-supervised) [75]	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
CLAN (Adversarial) [44]	31.46	65.39	21.14	69.10	17.29	25.49	11.17	3.14	7.61	71.03	28.19	55.55	18.86	2.76	71.60	26.42	17.99	59.53	9.44	15.91
CRST (Self-training) [97]	31.67	68.18	15.72	76.78	14.06	26.11	9.90	0.82	2.66	69.36	21.95	80.06	9.71	1.25	65.12	38.76	27.22	48.85	7.10	18.08
P2PDA (Adversarial) [88]	41.99	70.21	30.24	78.44	26.72	28.44	14.02	11.67	5.79	68.54	38.20	85.97	28.14	0.00	70.36	60.49	38.90	77.80	39.85	24.02
SIM (Self-training) [67]	44.58	68.16	32.59	80.58	25.68	31.38	23.60	19.39	14.09	72.65	26.41	87.88	41.74	16.09	73.56	47.08	42.81	56.35	47.72	39.30
PCS (Self-training) [84]	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
USSS (IDD) [30]	26.98	68.85	5.41	67.39	15.10	21.79	13.18	0.12	7.73	70.27	8.84	85.53	22.05	1.71	58.69	16.41	12.01	0.00	23.58	13.90
USSS (Mapillary) [30]	30.87	71.01	31.85	76.79	12.13	23.61	11.93	3.23	10.15	73.11	31.24	89.59	16.05	3.86	65.27	24.46	18.72	0.00	9.08	14.48
Seamless (Mapillary) [50]	34.14	59.26	24.48	77.35	12.82	30.91	12.63	15.89	17.73	75.61	33.30	87.30	19.69	4.59	63.94	25.81	57.16	0.00	11.59	19.04
SwiftNet (Cityscapes) [48]	25.67	50.73	32.76	70.24	12.63	24.02	18.79	7.18	4.01	64.93	23.70	84.29	14.91	0.97	43.46	8.92	0.04	4.45	12.77	8.77
SwiftNet (Merge3) [90]	32.04	68.31	38.59	81.48	15.65	23.91	20.74	5.95	0.00	70.64	25.09	90.93	32.66	0.00	66.91	42.30	5.97	0.07	6.85	12.66
Trans4PASS-S (ours)	**55.25**	78.39	41.62	**86.47**	31.56	45.47	34.02	22.98	18.33	79.63	41.35	93.80	49.02	22.99	81.05	67.43	**69.64**	86.04	60.85	39.20
Trans4PASS-S (ours)*	**56.38**	79.91	42.68	86.26	30.68	42.32	**36.61**	**24.81**	**19.64**	78.80	**44.73**	93.84	**50.71**	**24.39**	81.72	**68.86**	66.18	**88.62**	**63.87**	**46.62**

(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.

Network	Method	mIoU(%)
FANet	P2PDA	35.67
DANet	P2PDA	41.99
Trans4PASS-T	P2PDA	51.05
Trans4PASS-S	P2PDA	52.91
Trans4PASS-T	-	45.89
Trans4PASS-T	Warm-up	50.56
Trans4PASS-T	SSL	51.86
Trans4PASS-T	MPA	51.93
Trans4PASS-T	MPA + SSL	53.26
Trans4PASS-T	MPA + SSL + MS	**54.72**
Trans4PASS-S	-	48.73
Trans4PASS-S	Warm-up	52.59
Trans4PASS-S	SSL	54.67
Trans4PASS-S	MPA	54.77
Trans4PASS-S	MPA + SSL	55.25
Trans4PASS-S	MPA + SSL + MS	**56.38**

(b) **Adaptation results on DensePASS.**

Network	Method	mIoU(%)
DANet	-	40.28
DANet	P2PDA	42.26
PVT-Tiny	-	24.45
PVT-Tiny	P2PDA	39.66
PVT-Small	-	23.11
PVT-Small	P2PDA	43.10
Trans4PASS-T	-	46.08
Trans4PASS-T	MPA	47.48
Trans4PASS-S	-	48.34
Trans4PASS-S	MPA	**52.15**
DANet	Supervised	44.15
Trans4PASS-S	Supervised	**53.31**

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

	Method	Input	mIoU(%)
Supervised	StdConv [59]	RGB	32.6
	CubeMap [59]	RGB	33.8
	DistConv [59]	RGB	34.6
	UNet [53]	RGB-D	35.9
	GaugeNet [10]	RGB-D	39.4
	UGSCNN [28]	RGB-D	38.3
	HexRUNet [86]	RGB-D	43.3
	Tangent [16] (ResNet-101)	RGB	45.6
	HoHoNet [58] (ResNet-101)	RGB	52.0
	Trans4PASS (Small)	RGB	52.1
	Trans4PASS (Small+MS)	RGB	**53.0**
UDA	Trans4PASS (Source only)	RGB	48.1
	Trans4PASS (MPA)	RGB	50.8
	Trans4PASS (MPA+MS)	RGB	**51.2**

(d) **Comparison on SPan** averaged by 3 folds.

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

4.4. PIN2PAN Adaptation

Ablations in outdoor scenarios. To verify the generalization ability of applying Trans4PASS in adaptation methods, FANet and DANet used in P2PDA [45] are replaced by Trans4PASS-T/-S, as visible in Table 4b. Trans4PASS 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 Trans4PASS-T and -S models, surpassing 51.05% and 52.91% of P2PDA. The second and third ablation groups of Table 4b show how Trans4PASS-T and -S match up against each other. Individually, MPA is on par with the SSL-based method. When combining both, MPA and SSL, Trans4Pass-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.

Omnidiretional segmentation. To showcase the effectiveness of MPA on omnidiretional segmentation, the panoramic image is divided into 8 directions and evaluated individually. The polar diagram in Fig. 5 demonstrates that MPA brings uniform improvement to omnidirectional segmentation.

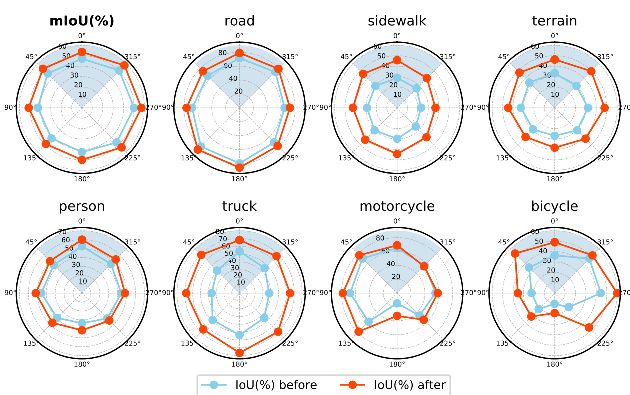
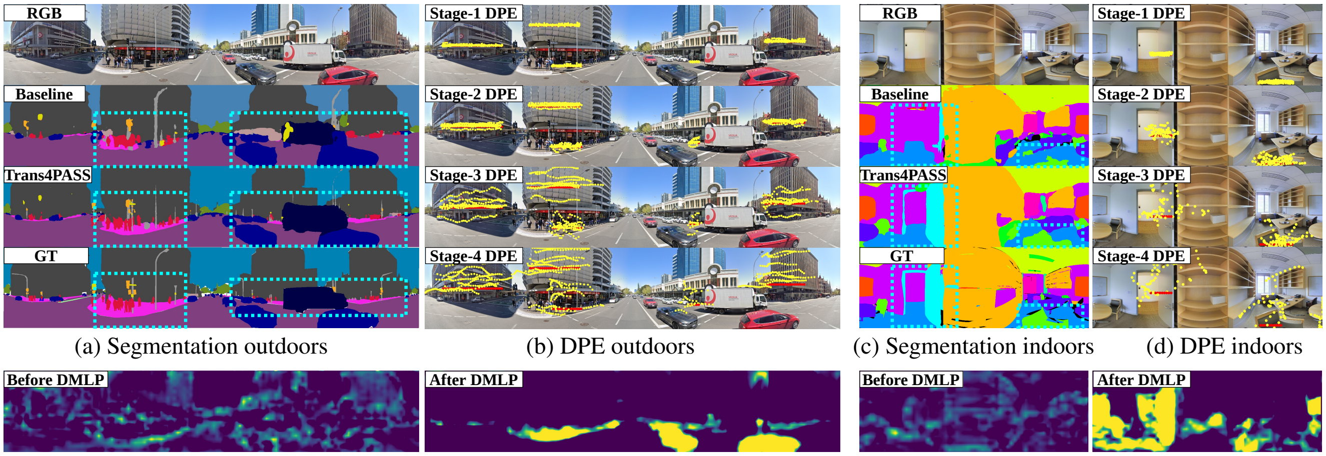


Figure 5: **Comparison of omnidirectional segmentation** before and after mutual prototypical adaptation.

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°, IoUs 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,



(a) Segmentation outdoors (b) DPE outdoors (c) Segmentation indoors (d) DPE indoors

(e) DMLP outdoors (f) DMLP indoors

Figure 6: **Qualitative comparisons, DPE and DMLP visualizations.** (a) and (c) show segmentation comparisons, where the baseline has neither DPE/DMLP nor MPA. The ● dots in (b) and (d) are sampling points shifted by learned offsets *w.r.t.* the ● patch center of DPE (from decoder). (e) and (f) show the #75 channel maps of stage-3 before and after DMLP. Zoom in for better view.

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 ($\sim$14M parameters) obtains a high mIoU score (51.2%), even comparable to existing fully-supervised and transfer-learning methods, which are based on ResNet-101 backbones ($\sim$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 75^{th} 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

[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, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao 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

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

[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 surround view cameras. *T-ITS*, 2020. 2

[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

[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

[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] 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 *ICCV*, 2021. 2

[20] Julia Guerrero-Viu, 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

[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

[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:2107.08391*, 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, Zihang 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-SfMLearner: Self-Supervised multi-task learning of depth and semantics in panoramic videos. *SPL*, 2021. 2

[41] Yahao Liu, Jinhong Deng, Xinchen 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, Yixuan 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 adversaries for semantics 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

[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. *SIVP*, 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

[54] Ahmed Rida Sekkat, Yohan Dupuis, Pascal Vasseur, and Paul Honeine. The OmniScape dataset. In *ICRA*, 2020. 2

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

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

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

[58] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021. 1, 2, 7

[59] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, 2018. 1, 7

[60] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021. 2, 4

[61] 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

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

[63] 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

[64] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[65] 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 *ICCV*, 2021. 3, 4, 6, 8

[66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[67] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020. 7, 8

[68] 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 *NeurIPS*, 2021. 3, 4, 6, 8

[69] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *TPAMI*, 2019. 1

[70] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. Semantic segmentation of panoramic images using a synthetic dataset. In *SPIE*, 2019. 2

[71] Yanyu Xu, Ziheng Zhang, and Shenghua Gao. Spherical DNNs and their applications in 360° images and videos. *TPAMI*, 2021. 1

[72] Kailun Yang, Xinxin Hu, Luis Miguel Bergasa, Eduardo Romera, and Kaiwei Wang. PASS: Panoramic annular semantic segmentation. *T-ITS*, 2020. 1, 7

[73] 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 *IV*, 2020. 1

[74] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild. *TIP*, 2021. 2

[75] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *CVPR*, 2021. 1, 2, 4, 7

[76] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 2

[77] Yaozu Ye, Kailun Yang, Kaite Xiang, Juan Wang, and Kaiwei Wang. Universal semantic segmentation for fisheye urban driving images. In *SMC*, 2020. 2

[78] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, 2020. 6

[79] Senthil Kumar Yogamani, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek O'Dea, Patrick Pérez, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricár, Stefan Milz, Martin Simon, and Karl Amende. WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *ICCV*, 2019. 2

[80] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR*, 2020. 2

[81] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 6

[82] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution transformer for dense prediction. In *NeurIPS*, 2021. 2

[83] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip H. S. Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *ICCV*, 2021. 4

[84] 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 *CVPR*, 2021. 2, 3, 4, 7, 8

[85] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. DeepPanoContext: Panoramic 3D scene understanding with holistic scene context graph and relation-based optimization. In *ICCV*, 2021. 2

[86] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *ICCV*, 2019. 7

[87] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola. ResNeSt: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 6

[88] 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. *T-ITS*, 2021. 2, 7, 8

[89] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In *ICCVW*, 2021. 4, 6

[90] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. ISSAFE: Improving semantic segmentation in accidents by fusing event-based data. In *IROS*, 2020. 7, 8

[91] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 2021. 3, 4

[92] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 2

[93] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 6

[94] 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 *CVPR*, 2021. 1, 2, 3, 4, 6

[95] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 2021. 2

[96] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 4

[97] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 2, 7, 8