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\cite{13,75} and bringing immersive viewing experiences to augmented- and virtual reality displays\cite{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\cite{13,20}. In particular, dense semantic segmentation on panoramic images, facilitates a high-level holistic pixel-wise understanding of surrounding environments\cite{45,73}.

Panoramic semantic segmentation is usually performed on 2D panoramas that were transformed using equirectangular projection\cite{45,75}, 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\cite{72} and as such they have to be adequately addressed. Considering the intricacies of panoramas, convolution variants\cite{10,55,59} and attention-augmented models\cite{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\cite{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 fish-eye 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 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 encompases 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 $HW \times 3$, 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 $f \in \mathbb{R}^{HW \times C_{in}}$, a standard PE module [15, 68] splits it into a flattened 2D patch sequence $z \in \mathbb{R}^{(\frac{HW}{s^2}) \times (s^2 \cdot C_{in})}$, 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 $z$ 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 $\Delta(i, j) \in [\frac{-s}{2}, \frac{s}{2}]^2$. 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 $\Delta_{DPE} \in \mathbb{N}^{HW \times 2 \times 2}$ 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).

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

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 imagery 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 profiling 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:

\[
\hat{z}(k,c) = \sum_{k=1}^{HW} \sum_{c=1}^{C_{in}} w(k,c) \cdot z \in \mathbb{R}^{HW \times C_{in}},
\]

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:

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

where \( \text{Up}(\cdot) \) and \( \text{LN}(\cdot) \) 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 $\mathcal{D}_s^x = \{ (x_i^s, y_i^s) \}_{x^s \in \mathbb{R}^H \times W \times 3}, y^s \in \{0, 1\}$ and the target (panoramic) dataset $\mathcal{D}_t^x = \{ (x_i^t) \}_{x^t \in \mathbb{R}^H \times W \times 3}$ 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:

$$\mathcal{L}_{\text{SEG}}^x = - \sum_{i,j,k} y_{i,j,k} \log(p_{i,j,k}^s)$$

where $p_{i,j,k}^s$ indicates the probability of pixel $x_{i,j}^s$ 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 $\hat{y}_{i,j,k}$ of pixels $x_{i,j}^s$ in the target domain:

$$\mathcal{L}_{\text{SSL}}^t = - \sum_{i,j,k} \hat{y}_{i,j,k} \log(p_{i,j,k}^t),$$

where the pseudo label is given by the most probable class in the model predictions: $\hat{y}_{i,j,k} = \mathbb{1}_{k = \arg \max \phi(x_{i,j}^s)}$. 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 $\mathcal{F} = \{ f_1, \ldots, f_{n_s} \} \cup \{ f_1^{\prime}, \ldots, f_{n_t}^{\prime} \}$, with feature maps $f$ fused from four-stage multi-scale features $f = \sum_{l=1}^4 f_l$. Each feature map is associated either with its respective source ground-truth label or a target pseudo-label. To compute the mutual prototype memory $\mathcal{M} = \{ P_1, \ldots, P_K \}$ with prototypes $P_k$ we take the mean of all feature vectors (pixel-embeddings) from all feature maps in $\mathcal{F}$ 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 $P_{k}^t$ 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 $f$ to be aligned with the prototypical feature map $\tilde{f}$ 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 $\tilde{f}$ has the same shape as $f$. For brevity, only the source domain is displayed in Eq. (6), which is similar to the target domain.

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

where KL(·), 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:

$$\mathcal{L} = \mathcal{L}_{\text{SEG}}^s + \alpha (\mathcal{L}_{\text{MPA}}^s + \mathcal{L}_{\text{MPA}}^t).$$

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 folders, 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...
### Performance gaps of CNN- and transformer-based models from Cityscapes (CS) \( \times 1024 \times 512 \) to DensePASS (DP).

<table>
<thead>
<tr>
<th>Network</th>
<th>Backbone</th>
<th>CS</th>
<th>DP</th>
<th>mIoU Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwiftNet [48]</td>
<td>ResNet-18</td>
<td>75.4</td>
<td>25.7</td>
<td>-49.7</td>
</tr>
<tr>
<td>Fast-SCNN [51]</td>
<td>Fast-SCNN</td>
<td>69.1</td>
<td>24.6</td>
<td>-44.5</td>
</tr>
<tr>
<td>ERFNet [52]</td>
<td>ERFNet</td>
<td>72.1</td>
<td>16.7</td>
<td>-55.4</td>
</tr>
<tr>
<td>FAaN [24]</td>
<td>ResNet-34</td>
<td>71.3</td>
<td>26.9</td>
<td>-44.4</td>
</tr>
<tr>
<td>PSPNet [93]</td>
<td>ResNet-50</td>
<td>78.6</td>
<td>29.5</td>
<td>-49.1</td>
</tr>
<tr>
<td>OCRNet [81]</td>
<td>HRNetV2p-W18</td>
<td>78.6</td>
<td>30.3</td>
<td>-47.8</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>ResNet-101</td>
<td>80.9</td>
<td>32.5</td>
<td>-48.4</td>
</tr>
<tr>
<td>DANet [17]</td>
<td>ResNet-101</td>
<td>80.4</td>
<td>28.5</td>
<td>-51.9</td>
</tr>
<tr>
<td>DNL [78]</td>
<td>ResNet-101</td>
<td>80.4</td>
<td>32.1</td>
<td>-48.3</td>
</tr>
<tr>
<td>Semantic-FPN</td>
<td>ResNet-101</td>
<td>75.8</td>
<td>28.8</td>
<td>-47.0</td>
</tr>
<tr>
<td>ResNet [87]</td>
<td>ResNet-101</td>
<td>79.6</td>
<td>28.8</td>
<td>-50.8</td>
</tr>
<tr>
<td>OCRNet [81]</td>
<td>HRNetV2p-W48</td>
<td>80.7</td>
<td>32.8</td>
<td>-47.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th>Backbone</th>
<th>CS</th>
<th>DP</th>
<th>mIoU Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>SETR-Naive [94]</td>
<td>Transformer-L</td>
<td>77.9</td>
<td>36.1</td>
<td>-41.8</td>
</tr>
<tr>
<td>SETR-MLA [94]</td>
<td>Transformer-L</td>
<td>77.2</td>
<td>35.6</td>
<td>-41.6</td>
</tr>
<tr>
<td>SETR-PUP [94]</td>
<td>Transformer-L</td>
<td>79.3</td>
<td>35.7</td>
<td>-43.6</td>
</tr>
<tr>
<td>SegFormer-B1 [68]</td>
<td>SegFormer-B1</td>
<td>78.5</td>
<td>38.5</td>
<td>-40.0</td>
</tr>
<tr>
<td>SegFormer-B2 [68]</td>
<td>SegFormer-B2</td>
<td>81.0</td>
<td>42.4</td>
<td>-38.6</td>
</tr>
<tr>
<td>Trans4PASS-T</td>
<td>Trans4PASS-T</td>
<td>79.1</td>
<td>41.5</td>
<td>-37.6</td>
</tr>
<tr>
<td>Trans4PASS-S</td>
<td>Trans4PASS-S</td>
<td>81.1</td>
<td>44.8</td>
<td>-36.3</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Network</th>
<th>Backbone</th>
<th>SPin</th>
<th>SPan</th>
<th>mIoU Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwiftNet [48]</td>
<td>ResNet-18</td>
<td>42.28</td>
<td>34.95</td>
<td>-7.87</td>
</tr>
<tr>
<td>DANet [17]</td>
<td>ResNet-50</td>
<td>43.33</td>
<td>37.76</td>
<td>-5.57</td>
</tr>
<tr>
<td>DANet [17]</td>
<td>ResNet-101</td>
<td>40.09</td>
<td>31.81</td>
<td>-8.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network</th>
<th>Backbone</th>
<th>SPin</th>
<th>SPan</th>
<th>mIoU Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans4PASS-T</td>
<td>PVT-T</td>
<td>41.28</td>
<td>24.45</td>
<td>-16.83</td>
</tr>
<tr>
<td>Trans4PASS-S</td>
<td>PVT-S</td>
<td>44.47</td>
<td>23.11</td>
<td>-21.36</td>
</tr>
<tr>
<td>Trans4PASS-T</td>
<td>Trans4PASS-T</td>
<td>49.05</td>
<td>46.08</td>
<td>-2.97</td>
</tr>
<tr>
<td>Trans4PASS-S</td>
<td>Trans4PASS-S</td>
<td>50.20</td>
<td>48.34</td>
<td>-1.86</td>
</tr>
</tbody>
</table>

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

### 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. Table 1 summarizes the results tested on Cityscapes and DensePASS validation sets. Although previous transformers [68, 94] reduce the mIoU gap from \( \sim 50\% \) of CNN-based counterparts to \( \sim 40\% \), 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 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.

### Network Encoder Decoder | GFLOPs | #P | CS | DP |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans4PASS</td>
<td>MiT-B1* DMLP</td>
<td>13.11 13.10 69.48 36.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans4PASS</td>
<td>MiT-B1 CycleMLP</td>
<td>9.83 13.60 73.49 40.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans4PASS</td>
<td>MiT-B1 ASMLP [36]</td>
<td>13.40 14.19 73.65 42.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans4PASS</td>
<td>MiT-B1 DMLP</td>
<td>12.02 13.93 72.49 45.89 (+9.39)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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) \( \times 512 \times 512 \) and tested on DensePASS (DP) \( \times 2048 \times 400 \).

---

1 MMSegmentation: https://github.com/open-mmlab/mmsegmentation.
MPA works collaboratively with pseudo labels and provides uniform improvement to omnidirectional segmentation. 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. 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 segmentations. 

Figure 5: Comparison of omnidirectional segmentation before and after mutual prototypical adaptation.
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 6f. 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 $360^\circ$ 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 $180^\circ$. Transferring models between pinhole-, fisheye-, and panoramic domains, fusing modalities, and solving various tasks of $360^\circ$ imagery are opportunities for further research.
References


[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


[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


[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


[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


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


[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


