Look at the Neighbor: Distortion-aware Unsupervised Domain Adaptation for Panoramic Semantic Segmentation

Xu Zheng\textsuperscript{1} Tianbo Pan\textsuperscript{1} Yunhao Luo\textsuperscript{3} Lin Wang\textsuperscript{1,2,}\textsuperscript{*}
\textsuperscript{1}AI Thrust, HKUST(GZ) \textsuperscript{2}Dept. of CSE, HKUST \textsuperscript{3}Brown University
zhengxu128@gmail.com, tpan695@connect.hkust-gz.edu.cn, devinluo27@gmail.com, linwang@ust.hk

Project Page: https://vlislab22.github.io/DATR/

Figure 1: (a) We propose a novel UDA method that effectively addresses the distortion problems via the distortion-aware attention (DA) module that extracts more salient textural details than those of self-attention (SA) \cite{45}; (b) Our method achieves more than 8\% of mIoU than the SOTA method (with 24.98M parameters) \cite{45} while taking only 4.64M parameters on the Synthetic \cite{46}-to-Real \cite{20} scenarios.

Abstract

Endeavors have been recently made to transfer knowledge from the labeled pinhole image domain to the unlabeled panoramic image domain via Unsupervised Domain Adaptation (UDA). The aim is to tackle the domain gaps caused by the style disparities and distortion problem from the non-uniformly distributed pixels of equirectangular projection (ERP). Previous works typically focus on transferring knowledge based on geometric priors with specially designed multi-branch network architectures. As a result, considerable computational costs are induced, and meanwhile, their generalization abilities are profoundly hindered by the variation of distortion among pixels. In this paper, we find that the pixels’ neighborhood regions of the ERP indeed introduce less distortion. Intuitively, we propose a novel UDA framework that can effectively address the distortion problems for panoramic semantic segmentation. In comparison, our method is simpler, easier to implement, and more computationally efficient. Specifically, we propose distortion-aware attention (DA) capturing the neighboring pixel distribution without using any geometric constraints. Moreover, we propose a class-wise feature aggregation (CFA) module to iteratively update the feature representations with a memory bank. As such, the feature similarity between two domains can be consistently optimized. Extensive experiments show that our method achieves new state-of-the-art performance while remarkably reducing 80\% parameters.

1. Introduction

The burgeoning demand for omnidirectional and dense scene understanding has stimulated the popularity of 360° cameras, which pose much wider field-of-view (FoV) in the range of 360° × 180° than the 2D images captured by pin-
hole cameras [1]. 360° cameras deliver complete scene details either in the outdoor or indoor environment; therefore, research has been actively focused on panoramic semantic segmentation for the pixel-wise scene understanding of the intelligent systems, such as self-driving and augmented/virtual reality [26, 36, 39].

Generally, 360° images are projected into the 2D planar representations while preserving the omnidirectional information [40, 14, 50], to be aligned with the existing pipelines [31, 49]. Equirectangular projection (ERP) is the most commonly used projection type. However, ERP images often suffer from the image distortion and object deformation [46], caused by the non-uniformly distributed pixels. Also, the lack of precisely annotated datasets heavily impedes training effective panoramic semantic segmentation models.

For these reasons, research endeavors have been made to transfer knowledge from the labeled pinhole image domain to the unlabeled panoramic image domain via Unsupervised Domain Adaptation (UDA). It aims to tackle the domain gaps caused by intrinsic style disparities and inevitable distortion problems. Typically, [43, 44, 20, 9, 46, 45, 28] leverage the spatial geometric priors (e.g., convolution variants [28] and attention-augmented components [45, 46]) to address the distortion problems. However, these priors are essentially inadequate for the panoramic semantic segmentation; therefore, cumbersome, i.e., multi-branch network architectures [38] are designed to reinforce the learning abilities. Consequently, considerable computation costs are induced, and their generalization abilities are profoundly plagued by the variation of distortion among the pixels.

In this paper, we find that the pixels’ neighboring regions in the ERP indeed introduce less distortion. As the ERP shuffles the equidistribution of spherical pixels, the distance (Fig. 2 (b)) between any two pixels for a specific latitude of a 360° image is different from that (Fig. 2 (c)) of the ERP image (sphere-to-plane projection). As a result, it is easier to capture the positional distribution among the pixels by reducing the receptive field, which is more efficient in addressing distortion problems. Therefore, controlling the neighboring region size is crucial in balancing the trade-off between receptive field and distortion problems.

In light of this, we propose a novel UDA framework that can efficiently address the distortion problems for panoramic semantic segmentation. Compared with the state-of-the-art UDA methods [13, 45, 46, 41, 44], our method is simpler, easier to implement, and more computationally efficient. Our method enjoys two key contributions. Firstly, we propose a novel distortion-aware attention (DA) module to capture the neighboring pixel distributions between domains (See Fig. 1 (a)). This is buttressed by a trainable relative positional encoding (RPE), which provides unique neighboring positional information. We then build a hierarchically structured DA-based transformer (DATR) that aggregates the feature information from all layers. In addition, we propose a class-wise feature aggregation (CFA) module that transfers knowledge of the extracted features between domains. It updates the class-wise feature centers with a memory bank and consistently optimizes the cross-domain feature similarity by iteratively updating class centers.

We conduct extensive experiments for both the synthetic and real-world scenarios, including Cityscapes-to-DensePASS and Synthetic-to-DensePASS datasets. The results show that our framework surpasses the SOTA methods by +8.76% and +1.59% on the Synthetic-to-Real and Pinhole-to-Panoramic scenarios, respectively while taking only 20% parameters (See Fig. 1 (b)). In summary, our major contributions are three-fold: (I) Our work serves as the first attempt to address the distortion problems by capturing the neighboring pixel distributions. (II) We propose a DA module to capture the neighboring pixel distributions. (III) We propose a CFA module to iteratively transfer the cross-domain knowledge.

2. Related Work

UDA for Panoramic Semantic Segmentation can be divided into three major categories: adversarial learning, pseudo label generation, and feature prototype adaption. The first type of methods [11, 4, 25, 29] tends to learn the domain invariance by conducting alignment from the image level [11, 16, 22], feature level [11, 2, 12], and output level [19, 21]. The second type of methods generates pseudo labels for the target domain training and utilizes self-training to refine them. For example, [17, 42, 32, 47] conduct refinement by leveraging the guidance from an auxiliary task, e.g., depth estimation [17]. The third type of method, e.g., Mutual Prototype Adaption (MPA) [45], aligns the feature embeddings with the prototypes obtained in the source and target domain individually. However, these approaches utilize the multi-stage training strategy and thus fail to correlate the features in each mini-batch. Differently, we propose the CFA module to aggregate the class-wise prototypes and iteratively update them, promoting the prototypes to have a more holistic representation of the peculiarity of domains.

Distortion Problems of ERP. Previous works on alleviating the distortion problems are with the manner of deformable kernel [27, 6] and designing adaptable CNN [28, 48] according to the geometrical priors of the sphere. Particularly, [46] adaptively adjusts the receptive field during patch embedding to better retain the semantic consistency and considers the distortion problems during the feature parsing process. However, this method is inevitably inefficient due to the large receptive field (See
Another way is designing distortion-aware neural networks [45, 46]. Deformable components, e.g., Deformable Patch Embedding (DPE) and Deformable MLP (DMLP) [45], are widely explored for panoramic semantic segmentation as they can help to learn the prior knowledge of panorama characteristics when patchifying the input data. Though the performance is significantly improved, the abilities of data generalization are limited, and mostly rely on prior geometry knowledge. Differently, we find that the neighboring region of ERP indeed introduces less distortion, benefiting the generalization to the variance of pixel distribution. Therefore, we propose the DA module to address the distortion problem with much fewer parameters.

Self-attention (SA) is defined as a dot product operation on query, key, and value sequence [30]. Dosovitskiy et al. [8] first proposed to utilize the SA on image patches in the vision field. More recently, abundant variants of attention paradigms [18, 35, 23, 5] are proposed to tackle vision problems. For panoramic semantic segmentation, Multi-Head Self-Attention (MHSA) [30] and Efficient Self-Attention (ESA) [35] are extensively applied to capture the long-range dependencies of 360° images. However, MHSA and ESA are inadequate in alleviating the distortion problem between pixels caused by the global feature extraction strategy. By contrast, our work shares a different spirit by focusing on the neighboring pixels, and accordingly the DA module is proposed to reduce the distortion problems by capturing the distinct pixels’ distribution between domains.

3. Method

3.1. Theoretical Analysis of ERP Distortion

The most notable advantage of ERP is its ease of operation due to its projection strategy, which is a completely linear transformation [1]. However, ERP shuffles the equidistribution of sphere pixels, resulting in varying distances between pixels at different latitudes before and after ERP. As shown in Fig. 2 (a), (b), and (c), where the distances between pixels $w_1$, $w_2$, and $w_0$ are different.

$$w_0 = W/n, w_1 = \frac{2\pi}{n} \sqrt{h_1 \left(\frac{W}{\pi} - h_1\right)},$$
$$w_2 = \frac{2\pi}{n} \sqrt{h_2 \left(\frac{W}{\pi} - h_2\right)}, w_0 > w_2 > w_1,$$

where $n$ is the number of sampling pixels at each latitude, $W$ is the width of ERP and $h_1$, $h_2$ is the height of red and blue pixels. The non-uniform sampling density of spherical data caused by the ERP projection can lead to distortion, as the sampling density is different at the poles and equator.

Our idea: We quantify the lateral distortion as the difference in distance between the pixels on a distinct projection type. By contrast, the vertical distortion is evenly distributed. Intuitively, we formulate the distortion coefficient $Dis$ between the red and green points in Fig. 2 as:

$$Dis = w_4 - w_3 = \frac{n'}{n} (W - 2\pi \sqrt{h_0 \left(\frac{W}{\pi} - h_0\right)}),$$

where $n'$ is the number of pixels between the red and green points in the latitude coordinate. Eq.2 demonstrates that $Dis$ increases with $n'$, indicating that a smaller value of $n'$ leads to less distortion. This observation leads us to propose addressing the distortion problem in UDA for panoramic semantic segmentation from a neighboring perspective. By minimizing the receptive field of the UDA network model, we can more effectively capture the pixel’s distribution and address the distortion problem. Based on this theoretical analysis, we propose a distortion-aware attention (DA) module in Sec.3.2.1 to mitigate this issue. More detailed analysis see supplmat.

3.2. Distortion-aware UDA Framework

3.2.1 Distortion-aware Attention (DA)

We now describe how to tackle the distortion problem with the proposed DA module. It captures the pixel distribution
in the neighborhood region. As shown in Fig. 1(a), the pixels in the DA only look at the neighboring pixels. To get the intuition, let’s first review the MHSA [30], which can be formulated as:

\[
\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_{\text{head}}}})V, \tag{3}
\]

where \(Q\), \(K\), and \(V\) are the linear projections of the whole input sequence. Based on our observation, we propose to learn UDA networks efficiently without strong geometric constraints on the pixel’s neighboring regions. To achieve this, we narrow the receptive field and focus more on the local distribution of neighboring pixels to minimize the distortion coefficient, as defined in Eq. 2. To be specific, we denote the nearest pixels of \(p_{i,j}\) in the \(H \times W\) neighborhood region as \(P_{i,j}\), where \(H\) and \(W\) represent the height and width of the region, respectively. Thus, the DA can be defined as:

\[
DA(p_{i,j}) = \text{Softmax}(\frac{Qp_{i,j}K_{P_{i,j}}^T}{\text{scale}})V_{P_{i,j}}, \tag{4}
\]

where \(Q\), \(K\), and \(V\) are the linear projections of the corresponding input, the neighboring pixels, and have the same dimensions \(N \times C\), and \(N = H \times W\).

**Relative Positional Encoding (RPE).** Self-attention-based models suffer from the inability to capture the order of input tokens or patches [34]. To address this limitation for panoramic semantic segmentation, it is crucial to incorporate explicit representations of positional information. Previous MHSA-based methods [45, 46] leverage the geometric properties of the ERP to design specific fixed positional encoding. However, these methods rely heavily on given priors and lack adaptive distortion-aware abilities. In contrast, our RPE provides local neighboring positional information to alleviate this problem. As a result, our DA module can ignore the resolution difference between training and inference. As depicted in Fig. 3, it is evident that the distribution of neighboring pixels in the pinhole image (in blue) differs from that in the panoramic image (in yellow). To address this issue, we propose a trainable Relative Positional Encoding (RPE) for DA. Specifically, RPE consists of a set of trainable embedding vectors, initialized with a uniform distribution, that represents the position encoding of each pixel. The RPE captures the distorted pixel distribution of ERP after adaptation and remains fixed during inference. This encoding method captures the distribution of different neighboring pixels and enables the UDA model to address the domain gap resulting from the inevitable distortion. The encoding vectors are embedded into the DA module, allowing us to reformulate Eq.4 as follows:

\[
DA(p_{i,j}) = \text{Softmax}(\frac{Qp_{i,j}K_{P_{i,j}}^T}{\text{scale}})(V_{P_{i,j}} + \text{RPE}), \tag{5}
\]

where the RPE is our proposed positional encoding.

### 3.2.2 DA Block Design

We now present the design of the DA block for distortion-aware feature extraction. In contrast to prior methods that use a transformer backbone [45, 46], our DA block incorporates local position information through RPE, rather than a feed-forward network. As depicted in Figure 4(b), our DA block consists of normalization and MLP layers, making the feature extraction process completely convolution-free. Additional details of the DA block are in the supplmat.
3.2.3 DA-based Transformer

With the proposed DA block, we introduce a DA-based transformer model that is partially inspired by ViT [8]. Differently, our DA-based transformer model is tailored to address distortion problems and serves as a backbone for UDA in panoramic semantic segmentation. Our model adopts an encoder-decoder structure without extra components, resulting in fewer parameters than multi-branch models such as PASTS [15] (4.64M vs. 614M). In experiments, our DATR outperforms existing complex models [35, 45, 13, 46, 33] (see Tab.2) and is more robust to adaptation stability and efficiency (see Tab.3). More specifically, our model shares a different spirit from prior methods as it is tailored to specifically address the distortion problem in panoramic segmentation. Further details of our DA-based transformer model can be found in the supplmat.

**Encoder**. As shown in Fig. 4 (b), our DA-based transformer backbone can generate multi-scale features, including coarse shallow layers and fine-grained deep layers. To preserve the local continuity around the neighboring pixels, we use the overlapping patch merging of [35]. The patch size $K$, stride $S$, and padding size $P$ are set to 7/4/3 and 3/2/1 for the model with different scales. Concretely, given an input image with a resolution of $H \times W$, the hierarchical feature maps $F_i$ have the resolutions of $\frac{H}{2^i} \times \frac{W}{2^i}$, where $i \in \{1, 2, 3, 4\}$. For a better trade-off between receptive fields and distortion-aware abilities (See Tab. 4), the first three layers of our model are based on efficient self-attention, and the deepest layer is based on our DA block. To reduce the computational cost of ESA ($O(N^2)$), we adopt the sequence reduction process utilized in [33, 35].

**Decoder**. As depicted in Fig. 4 (b), the MLP decoder takes multi-scale features $F_i$ from the encoder as inputs, and the channel dimensions are aligned. Then $F_i$ are up-sampled to $F_1$’s size and concatenated and fused together. Finally, the last MLP layer takes the fused features to predict the segmentation confidence maps.

3.3. Class-wise Feature Aggregation (CFA)

Given the source (i.e., synthetic or pinhole images) domain dataset with a set of annotated images $X_s = \{x^s_i, y^s_i\}$, $x^s_i \in R^{H \times W \times 3}$, $y^s_i \in 0, 1^{H \times W \times K}$ and the target (i.e., panoramic images) dataset $D_t = \{x^t_i, y^t_i\}$, $x^t_i \in R^{H \times W \times 3}$ without corresponding labels, the objective of UDA is to transfer knowledge from the source domain to the target domain with $K$ shared classes. Our model is first trained with the source domain data $D_s$ using the segmentation loss:

$$\mathcal{L}_{SEG} = - \sum_{h,w,k=1}^{H,W,K} y^s_{(h,w,k)} \log(p^s_{h,w,k}),$$  

where $p^s_{(i,j,k)}$ is the prediction of the source image pixel $x^s_{(i,j)}$ as the $k$-th class. To adopt our model to the target domain data, we utilize the predictions of target images as the pseudo labels and perform self-supervised (SS) training. The calculation and generation metric is:

$$\hat{y}^t_{(h,w,k)} = 1_{k=\arg\max(p^t_{h,w,\cdot})}.$$  

With the pseudo labels, our model is optimized by:

$$\mathcal{L}_{SS} = - \sum_{h,w,k=1}^{H,W,K} \hat{y}^t_{(h,w,k)} \log(p^t_{h,w,k}).$$

We propose a novel class-wise feature aggregation (CFA) module to transfer knowledge from extracted features between domains, as illustrated in Fig.4 (a). Our CFA module differs from prior UDA methods [45, 46] in two distinct ways: (a) it iteratively aggregates class-wise features and updates feature centers, thereby aligning them directly between the two domains; (b) hard pseudo-labels $\hat{y}^t_{(h,w,k)}$ are softened in the feature space and used for prediction, enabling full leveraging of knowledge from the source domain. Specifically, we utilize the pseudo labels $\hat{y}^t_{(h,w,k)}$ as the class-wise masks in the high-level feature space. Given the target domain feature’s $i$th class center $T_i^t$ in $t$th iteration and the class center from the source domain feature is $S_i^t$, the extracted features are first masked by the pseudo labels and then are projected to class-wise feature centers $S_i^t \setminus T_i^t$. Note that $S_i^t \setminus T_i^t$ is mixed with the corresponding feature center $S_i^{t-1} \setminus T_i^{t-1}$ from the last iteration (mini-batch). We leverage this iterative mixing strategy for the feature center to make the centers (a.k.a, prototypes) more robust:

$$C_i^s = (1 - \frac{1}{e})S_i^{(t-1)} + \frac{1}{e}S_i^t, C_i^t = (1 - \frac{1}{e})T_i^{(t-1)} + \frac{1}{e}T_i^t,$$

where the $C_i^s$ and $C_i^t$ are the class centers of the source and target images, and $e$ is the current epoch number. The same class-wise feature centers are pushed together by the Mean Squared Error (MSE) to mimic the class-wise knowledge between domains:

$$\mathcal{L}_f = \frac{1}{num} \sum_{i\in C} (C_i^s - C_i^t)^2,$$

where $num$ is the total category number.

4. Experiments

To evaluate the effectiveness of our proposed UDA framework, we conducted extensive experiments on
both real-world and synthetic datasets, which include Cityscapes [7], DensePASS [20], and SynPASS [46]. We designed three instance models of the DATR, namely DATR-Mini, DATR-Tiny, and DATR-Small, with the same architecture but different sizes. DATR-Mini is the most efficient model, with the least computation cost during training and testing (4.64M). On the other hand, DATR-Small is the most powerful model with the best performance.

### 4.1. Datasets and Implementation Details

**Cityscapes** [7] is a widely used dataset of urban street images captured from 50 different cities, with precise annotations of 19 categories. The official split includes 2975 images for training and 500 images for validation. In this paper, we use the official training set as the source data.

**SynPASS** [46] is a dataset of 9080 synthetic panoramic images annotated with 22 categories. The official training, validation, and test sets contain 5700, 1690, and 1690 images, respectively. We use the overlapping 13 classes between SynPASS and DensePASS datasets for training and testing.

**DensePASS** [20] is a 360-degree image dataset collected from 40 cities. The official training and test sets contain 2000 and 100 images, respectively, and are annotated with the same 19 classes as Cityscapes.

**Implementation Details.** We train all the models using 4 NVIDIA GPUs, with an initial learning rate of $5 \times 10^{-5}$. The learning rate is scheduled using a polynomial strategy with a power of 0.9. We use the AdamW optimizer with an epsilon of $1 \times 10^{-8}$ and weight decay of $1 \times 10^{-4}$. The resolutions and data augmentations for the training and test images are kept the same as those used in [45].

### 4.2. Experimental Results

We first evaluate our proposed framework on Cityscapes (pinhole)-to-DensePASS (panorama) datasets. Tab. 1 presents the mIoU results evaluated on the Cityscapes and DensePASS test set. Our DATR without any adaption module has the least performance drop (36.75%) compared with the SOTA segmentation methods [3, 35, 45, 46]. Importantly, our DATR-M is much more compact than the prior

<table>
<thead>
<tr>
<th>Network</th>
<th>Backbone</th>
<th>CS</th>
<th>DP</th>
<th>Gap</th>
<th>Param</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ResNet-101 [10]</td>
<td>80.89</td>
<td>32.69</td>
<td>48.20</td>
<td>58.75</td>
</tr>
<tr>
<td>Segformer [35]</td>
<td>MiT-B1</td>
<td>78.85</td>
<td>36.98</td>
<td>41.87</td>
<td>14.72</td>
</tr>
<tr>
<td></td>
<td>MiT-B2</td>
<td>81.32</td>
<td>42.23</td>
<td>40.09</td>
<td>25.76</td>
</tr>
<tr>
<td>Trans4PASS [45]</td>
<td>Trans4PASS-T</td>
<td>80.54</td>
<td>41.56</td>
<td>38.98</td>
<td>13.95</td>
</tr>
<tr>
<td></td>
<td>Trans4PASS-S</td>
<td>81.17</td>
<td>42.47</td>
<td>38.70</td>
<td>24.98</td>
</tr>
<tr>
<td>Trans4PASS+ [46]</td>
<td>Trans4PASS-T+</td>
<td>79.92</td>
<td>41.33</td>
<td>38.59</td>
<td>13.95</td>
</tr>
<tr>
<td></td>
<td>Trans4PASS-S+</td>
<td>81.76</td>
<td>42.35</td>
<td>39.41</td>
<td>24.98</td>
</tr>
<tr>
<td>DATR</td>
<td>DATR-M</td>
<td>75.23</td>
<td>38.48</td>
<td><strong>36.75</strong></td>
<td>4.64</td>
</tr>
<tr>
<td></td>
<td>DATR-T</td>
<td>79.01</td>
<td>42.22</td>
<td><strong>36.79</strong></td>
<td>14.72</td>
</tr>
<tr>
<td></td>
<td>DATR-S</td>
<td>79.98</td>
<td>47.55</td>
<td><strong>32.43</strong></td>
<td>25.76</td>
</tr>
</tbody>
</table>

Table 1: Performance gaps of some SOTA CNN-based and transformer-based panoramic semantic segmentation models. The test size on DensePASS (DP) is 400 × 2048. All models are only trained with the Cityscapes dataset without the adaption module.
models, with only 4.64M parameters, and has a competitive and even superior performance. In Tab. 2, we compare DATR against previous SOTA approaches, including PASS [37], Omni-sup [39], P2PDA [44], PCS [41] and Trans4PASS [45]. Among these methods, though DAFormer [13] serves as the SOTA transformer-based DA method, Trans4PASS achieves better results. Yet our DATR-S reaches the mIoU of 56.81% on the DensePASS test set while outperforming the DAFormer and Trans4PASS by mIoU increment of +2.14% and +1.59%, respectively. Evidently, large improvements (over 3%) have been obtained on sidewalk, wall, pole, etc., which are challenging yet pivotal categories in practical applications.

We further evaluate our method on SynPASS (synthetic)-to-DensePASS (real-world) datasets. Though synthetic panoramas' spatial distributions and content are closer to the real panoramic data, the previous methods struggle on several categories, e.g., traffic light and traffic sign, which strongly rely on texture cues. The main reasons are two folds: 1) the simulated traffic elements in the synthetic data do not present diverse textures and details, which aggravate thedomain gap; 2) prior UDA methods fail to capture the regional pixel-wise correspondence. As shown in Tab. 3, all of the variants of our DATR consistently outperform previous SOTA methods. Especially, our DATR-S achieves dramatically mIoU increment on the most challenging categories, including Fence (+23.04%), Pole (+11.75%), Tr.Light (+18.29%), Tr.Sign (+16.75%), Person (+8.52%), and Car (+6.75%). This is also demonstrated in Fig. 5. Overall, even the most compact model, DATR-M, achieves +5.75% mIoU increment than the prior SOTA method Trans4PASS+-S [46](24.98M) with only 4.64M parameters. This indicates that our DATR model brings largely enhanced domain adaptation performances in synthetic-to-real scenarios, buttressed by the proposed DA and CFA modules.

### 5. Ablation Study and Analysis

**Design Choices for DA and ESA.** We first demonstrate the superiority and necessity for the simultaneous usage of ESA and DA by conducting ablation experiments. We set different experiment settings, including DA at different structures. As we can see in Tab. 4, only using ESA suffers unsatisfactory performance. Meanwhile, using DA in the shallow layers also leads to poor performance due to the limited local receptive fields of the DA module. By contrast, utilizing DA at the deep layer brings the
largest performance gain (+2.03% ↑). It validates that using ESA and DA together in the proper sequence makes DATR achieves better distortion-aware abilities.

**Effectiveness of Relative Positional Encoding.** We now study the effectiveness of the proposed RPE in DA. We test three positional encoding strategies including absolute positional encoding (APE), relative positional encoding (RPE), and without any positional encoding (without PE). Our DATR-M achieves 41.56%, 51.04%, and 36.98% mIoU with APE, RPE and without PE, respectively. This indicates the significant effectiveness of our proposed RPE in conferring distortion-aware abilities to models.

**Ablation ofUDA Module.** To verify the effectiveness of our proposed UDA module, we conduct experiments on two benchmarks with different modules. As shown in Tab. 5, evidently, models trained with CFA achieve over 6% mIoU increment on Cityscapes-to-DensePASS than the models only trained with the source (Cityscapes) data. Moreover, our CFA gives more than 9% mIoU improvement to all variants of DATR, which indicates the effectiveness of our CFA and the structure superiority of DATR. As for SynPASS to DensePASS, our Self-Sup strategy and CFA achieve more than 10% mIoU increment.

**Generality of CFA.** In SynPASS-to-DensePASS scenario, our CFA achieves 48.33% and 44.97% and 48.96% mIoU with Trans4PASS-S, PVT-S and DAFormer. Compared with the MPA Tab. 3, our CFA achieves higher UDA performance by 3.04%, 4.07%, and 3.71% than MPA, respectively. The results confirm that CFA provides a better learning signal to diverse models.

**Effectiveness of DATR.** We evaluate the generalization capacity of our introduced DATR by applying it in other domain adaptation methods, e.g., MPA [45]. As shown in Tab. 3, our DATR consistently achieves superior performance than PVT [33] and Trans4PASS+ [46]. Specifically, our most efficient instance model DATR-M (4.64M) outperforms PVT-S (24.5M) and Trans4PASS+S (24.98M) by +7.34% and 2.95% mIoU increments, respectively. Evidently, large improvements have been made by our DATR-S model, which achieves +11.86% and +7.47% compared with PVT-S and Trans4PASS+S. This indicates the superior generalization capacity of our proposed DATR, facing domain shift behind 360° and pinhole imagery.

### 6. Discussion

**Significance of Neighboring Region Size.** The DA’s cover size has a significant impact on the performance of our proposed DATR. When the neighborhood region covers the entire input, DA behaves like self-attention.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Source 38.48</th>
<th>Cityscapes</th>
<th>&gt; DensePASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATR-M SS</td>
<td>Source 38.48</td>
<td>DATR-M SS</td>
<td>DATR-M SS</td>
</tr>
<tr>
<td>CFA</td>
<td>CFA 52.90 +14.42</td>
<td>CFA 54.60 +12.40</td>
<td>CFA 51.04 +18.03</td>
</tr>
<tr>
<td>DATR-T SS</td>
<td>Source 42.22</td>
<td>DATR-T SS</td>
<td>DATR-T SS</td>
</tr>
<tr>
<td>CFA</td>
<td>CFA 54.60 +12.40</td>
<td>CFA 53.23 +19.25</td>
<td></td>
</tr>
<tr>
<td>DATR-S SS</td>
<td>Source 47.55</td>
<td>DATR-S SS</td>
<td>DATR-S SS</td>
</tr>
<tr>
<td>CFA</td>
<td>CFA 56.81 +9.26</td>
<td>CFA 54.05 +17.31</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Ablation study of the proper location of our proposed DA block. *: DA block, ◦: ESA block.

| Cityscapes |-| DensePASS | SynPASS |-| DensePASS |
|------------|----------|-----------|----------|----------|
| Backbone Method mIoU | δ | Backbone Method mIoU | δ |
| DATR-M SS | Source 38.48 | DATR-M SS | DATR-M SS |
| CFA | CFA 52.90 +14.42 | CFA 54.60 +12.40 | CFA 51.04 +18.03 |
| DATR-T SS | Source 42.22 | DATR-T SS | DATR-T SS |
| CFA | CFA 54.60 +12.40 | CFA 53.23 +19.25 | |
| DATR-S SS | Source 47.55 | DATR-S SS | DATR-S SS |
| CFA | CFA 56.81 +9.26 | CFA 54.05 +17.31 | |

Table 5: Ablation study of different module combinations on two public benchmarks.

However, using a smaller neighborhood size results in simpler distortion but smaller receptive fields, which can be fatal for semantic segmentation tasks. We conducted experiments with different neighborhood region sizes ranging from 7 × 7 to 15 × 15 and found that much smaller or larger region sizes did not bring satisfactory results. To balance GFLOPs and performance, we choose the region size as 11 × 11.

**Performance of small objects.** Our proposed DATR demonstrates its capability of performing well on relatively small objects, such as traffic lights, through the use of DA blocks with narrowed receptive fields of attention. This is evidenced by the performance gains on such objects, with an increase of 18.29% as presented in Tab. 3. Moreover, DATR also shows reasonable performance gains on larger objects, such as the sky (+2.66%) and buildings (+2.80%).

**Rationality of DA and ESA.** To demonstrate the segmentation performance on panoramic images, we conduct ablation experiments of training a single layer of DA and ESA. As shown in Fig. 6(c), our proposed DA shows more salient textural details than ESA w.r.t. the visualization of the extracted features, showing the superiority of our DA.

**Effectiveness and rationality of CFA.** To demonstrate the effectiveness and rationality of our proposed CFA, we conducted ablation experiments using our proposed DATR model. The results in Tab.3 show that our CFA consistently outperforms the previous state-of-the-art method MPA [45].
With all variants of our proposed DATR model, our CFA achieved the highest mIoU, surpassing MPA by 1.29%, 1.12%, and 2.80% mIoU, respectively. Additionally, we used t-SNE visualization to compare the extracted features before and after incorporating our CFA on the DensePASS test set. As shown in Fig. 6 (b), CFA provides holistic class-wise feature aggregation, resulting in more closely tied features within the same class, such as the red and blue circles.

Iterative Strategy in CFA. To ablate the impact of our proposed iterative class-wise feature aggregation module, we substitute the iterative strategy with non-iterative and linear iterative methods to demonstrate the effectiveness of our iterative strategy. We conduct experiments with DATR-M by non-iterative and linear iterative strategy, the mIoU results are 50.43% and 50.53%, respectively. Our iterative update makes the class center more robust and holistic and achieves 51.04% mIoU with the same backbone.

Analysis of failure classes. Our DATR, which uses DA blocks in the deepest layer, performs well on relatively small objects, such as traffic lights (+18.29%) shown in Tab. 3. It also achieves reasonable performance gains on large objects, such as sky (+2.66%) and building (+2.80%). Consequently, the performance of each class is related to the neighboring size of DA, as shown in Fig. 6(a).

Unrestricted resolutions. Our DATR computes attention in a fixed neighborhood region, which is not largely affected by input sizes. We test different resolutions during inference, and the performance variation among different input sizes did not exceed 10%, while compared models [45, 46], fluctuate over 20% (See Tab.4 in the suppl.).

7. Conclusion and Future Work

In this paper, we found that the pixels’ neighborhood regions of ERP indeed introduce less distortion. Based on our observation, we proposed a novel distortion-aware attention (DA) module that focuses on capturing the neighboring pixel distribution, buttressed by a trainable relative positional encoding (RPE). We further built a unified backbone model for panoramic semantic segmentation. Moreover, we proposed a class-wise feature aggregation (CFA) module to iteratively update the features with a memory bank. As such, we consistently optimized the feature similarity between domains. Our proposed framework significantly outperformed the SOTA UDA methods with an order of magnitude fewer parameters.

Limitation and future work: In this paper, the UDA is performed on a single source dataset to a single target dataset. It would be worth exploring the multi-source domain adaptation for panoramic semantic segmentation. Moreover, we will explore how to utilize the other projections of the sphere data to facilitate knowledge transfer.

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


[47] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban
