SGAT4PASS: Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation

Xuewei Li[1,2], **Tao Wu**[1], **Zhongang Qi**[2], **Gaoang Wang**[3], **Ying Shan**[2] and **Xi Li**[1,4*]
[1]College of Computer Science and Technology, Zhejiang University
[2]ARC Lab, Tencent PCG
[3]Zhejiang University-University of Illinois at Urbana-Champaign Institute, Zhejiang University
[4]Zhejiang – Singapore Innovation and AI Joint Research Lab, Hangzhou
{xueweili, taowucs}@zju.edu.cn, zhongangqi@tencent.com, gaoangwang@intl.zju.edu.cn,
yingsshan@tencent.com, xilizju@zju.edu.cn

Abstract

As an important and challenging problem in computer vision, PAnoramic Semantic Segmentation (PASS) gives complete scene perception based on an ultra-wide angle of view. Usually, prevalent PASS methods with 2D panoramic image input focus on solving image distortions but lack consideration of the 3D properties of original 360° data. Therefore, their performance will drop a lot when inputting panoramic images with the 3D disturbance. To be more robust to 3D disturbance, we propose our Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation (SGAT4PASS), considering 3D spherical geometry knowledge. Specifically, a spherical geometry-aware framework is proposed for PASS. It includes three modules, i.e., spherical geometry-aware image projection, spherical deformable patch embedding, and a panorama-aware loss, which takes input images with 3D disturbance into account, adds a spherical geometry-aware constraint on the existing deformable patch embedding, and indicates the pixel density of original 360° data, respectively. Experimental results on Stanford2D3D Panoramic datasets show that SGAT4PASS significantly improves performance and robustness, with approximately a 2% increase in mIoU, and when small 3D disturbances occur in the data, the stability of our performance is improved by an order of magnitude. Our code and supplementary material are available at https://github.com/TencentARC/SGAT4PASS.

1 Introduction

There has been a growing trend of practical applications based on 360° cameras in recent years, including holistic sensing in autonomous vehicles [de La Garanderie *et al.*, 2018; Ma *et al.*, 2021; Gao *et al.*, 2022], immersive viewing in augmented reality and virtual reality devices [Xu *et al.*, 2018; Xu *et al.*, 2021; Ai *et al.*, 2022], etc. Panoramic images with an ultra-wide angle of view deliver complete scene perception in many real-world scenarios, thus drawing increasing attention in the research community in computer vision. Panoramic semantic segmentation (PASS) is essential for omnidirectional scene understanding, as it gives pixel-wise analysis for panoramic images and offers a dense prediction technical route acquiring 360° perception of surrounding scenes [Yang *et al.*, 2021a].

Most existing PASS approaches use equirectangular projection (ERP) [Sun *et al.*, 2021; Yang *et al.*, 2021b] to convert original 360° data to 2D panoramic images. However,

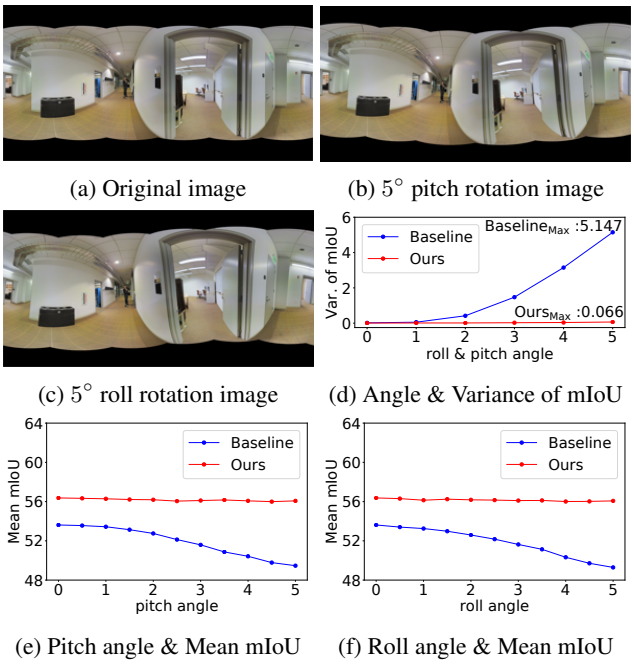


Figure 1: The results with 3D disturbance input. (a) is the original image, and (b) / (c) is the images rotated 5° in pitch / roll axis. Our baseline is Trans4PASS+. Compared with the minor change in images, the huge variance / performance change in SGA validation is shown in (d) / (e) and (f). "Mean" and "Variance" are defined in detail in Section 4.1.

*Corresponding author.

these methods often suffer from two main problems: large image distortions and lack of Spherical Geometry-Aware (SGA) robustness that resists 3D disturbance. These problems lead neural networks to only learn suboptimal solutions for panoramic segmentation [Yang *et al.*, 2019; Wang *et al.*, 2021a]. Although some recent works [Zhang *et al.*, 2022a; Zhang *et al.*, 2022b] take serious distortions into account in their models and become the current state-of-the-art (SOTA), they still do not pay enough attention to the SGA properties of the original 360° data, resulting in performance degradation even with small projection disturbance. As shown in Figure 1b and Figure 1c, applying 5° rotation on the pitch or roll axis of original 360° data carries only minor changes in 2D panoramic images. However, as shown in Figure 1e, Figure 1f, and Figure 1d, the performances of Trans4PASS+ [Zhang *et al.*, 2022b] (the blue lines) drop a lot (about 4%), and the variance increases by almost 2 orders of magnitude, because the axis rotations lead to different spherical geometry relations between pixels in the projected panoramic images, which the existing methods fail to adapt. Besides disturbance, the ERP also introduces boundaries to panoramic images that the original 360° data do not have. Some adjacent pixels are disconnected and some objects are separated, which is a severe issue, especially for semantic segmentation. Furthermore, there also exists a difference in pixel sampling density between the original 360° data and its corresponding projection image, e.g., pixels are over sampled in the antarctic and arctic areas of 2D panoramic images. All these issues make panoramic semantic segmentation a challenging task, and the above characteristics should be well studied to design a robust model that adapts to disturbance, disconnection, uneven density, and other SGA properties.

Improving robustness and taking SGA properties into account, we propose a novel model, i.e., Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation (SGAT4PASS), equipped with the SGA framework and SGA validation. The proposed SGA framework includes SGA image projection in the training process, Spherical Deformable Patch Embedding (SDPE), and a panorama-aware loss. SGA image projection provides images with 3D disturbance to improve the 3D robustness of the model. SDPE improves the patch embedding and makes it consider not only the image distortions with deformable operation but also spherical geometry with SGA intra- and inter-offset constraints. The panorama-aware loss deals with the difference in pixel density between the original 360° data and its corresponding 2D panoramic images. Moreover, we propose a new validation method, i.e., SGA validation, to evaluate the 3D robustness of various models comprehensively, which considers different 3D disturbances for input images, and measures the average performance and the variance for comparisons. Extensive experimental results on popular Stanford2D3D panoramic datasets [Armeni *et al.*, 2017] demonstrate that our proposed approach achieves about 2% and 6% improvements on traditional metrics and SGA metrics, respectively.

The contributions of this paper are summarized as follows:

- We propose SGAT4PASS, a robustness model for the PASS task, which utilizes SGA image projection to deal with the 3D disturbance issue caused by ERP.
- We introduce SDPE to combine spherical geometry with deformable operation to better deal with panoramic image distortion. And we also propose panorama-aware loss to ease the oversampling problem.
- We evaluate SGAT4PASS on the popular benchmark and perform extensive experiments with both traditional metrics and proposed SGA metrics, which demonstrate the effectiveness of each part of the framework.

2 Related Work

The two most related fields are panoramic semantic segmentation and dynamic and deformable vision transformers.

2.1 Panoramic Semantic Segmentation

Semantic segmentation of panoramic images has many applications in real-world scenarios, such as autonomous driving [Ye *et al.*, 2021], panoramic lenses safety and monitoring applications [Poulin-Girard and Thibault, 2012], etc. With the development of deep learning, many neural networks have been developed for panoramic semantic segmentation. Deng et al. [Deng *et al.*, 2017] first proposed a semantic segmentation framework for wide-angle (fish-eye) images and transformed an existing pinhole urban scene segmentation dataset into synthetic datasets. Yang et al. [Yang *et al.*, 2019] designed a semantic segmentation framework for panoramic annular images using a panoramic annular camera with an entire Field of View (FoV) for panoramic surrounding perception based on a single camera. Furthermore, Yang et al. [Yang *et al.*, 2020] proposed DS-PASS to improve it with a more efficient segmentation model with attention connections. PASS solutions can be divided into two main fields: distortion-aware strategies and 2D-geometry-aware ones.

For distortion-aware strategies, Tateno et al. [Tateno *et al.*, 2018] proposed using specially designed distortion-aware convolutions in a fixed manner to address image distortions. Furthermore, ACDNet [Zhuang *et al.*, 2022] combined convolution kernels with different dilation rates adaptively and used fusion-driven methods to take advantage of several projections. Jiang et al. [Jiang *et al.*, 2019] designed a spherical convolution operation. Lee et al. [Lee *et al.*, 2018] used spherical polyhedrons to represent panoramic views to minimize the difference in spatial resolution of the surface of the sphere and proposed new convolution and grouping methods for the representation of spherical polyhedrons. Hu et al. [Hu *et al.*, 2022] designed and proposed a distortion convolutional module based on the image principle to solve the distortion problem caused by the distortion of the panoramic image. Zhang et al. [Zhang *et al.*, 2022a] [Zhang *et al.*, 2022b] designed their Trans4PASS and Trans4PASS+ that perceived spherical distortion and solved the distortion problem of spherical images better through their Deformable Patch Embedding (DPE) and Deformable Multi-Layer Perception (DMLP) modules. Also, Trans4PASS+ is the current SOTA panoramic semantic segmentation model and is our baseline. For 2D geometry-aware strategies, horizontal features are mainly used based on the ERP inherent property. Sun et al. [Sun *et al.*, 2021] proposed HoHoNet and Pintore

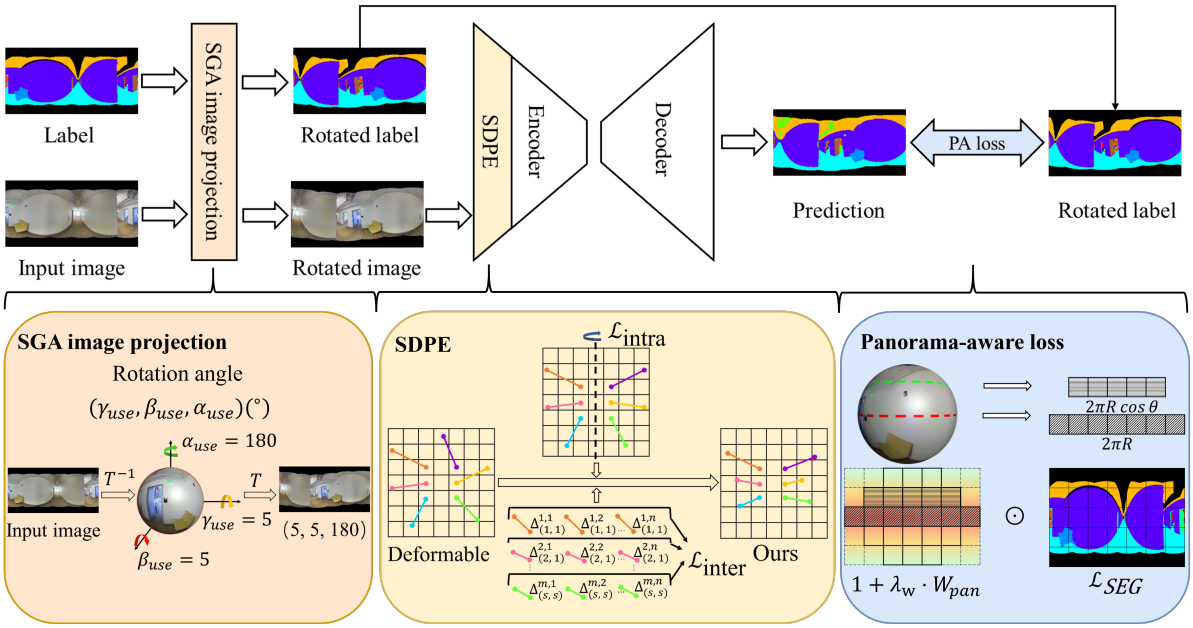


Figure 2: Overall review of SGAT4PASS. We borrow the network from Trans4PASS+, and add three main modules: Spherical geometry-aware (SGA) image projection, SDPE, and panorama-aware loss. (Lower left) SGA image projection rotates the input panoramic images to mimic 3D disturbance. (Lower middle) SDPE adds several SGA constraints on deformable patch embedding and let it consider both image distortions and spherical geometry. (Lower right) Panorama-aware loss (PA loss) takes into account the pixel density of a sphere.

et al. [Pintore *et al.*, 2021] proposed SliceNet to use the extracted feature maps in a 1D horizontal representation.

For our SGAT4PASS based on the distortion-aware SOTA model, Trans4PASS+, we add SGA information from the original 360° data instead of the 2D geometry prior to panoramic images to improve not only its performance but also its robustness when meeting 3D disturbance.

2.2 Dynamic and Deformable Vision Transformers

Regarding the field of vision transformers, some works have developed architectures with dynamic properties. Chen et al. [Chen *et al.*, 2021] and Xia et al. [Xia *et al.*, 2022] used deformable designs in later stages of the encoder. Yue et al. [Yue *et al.*, 2021] used a progressive sampling strategy to locate discriminatory regions. Deformable DETR [Zhu *et al.*, 2020] used deformable attention to deal with feature maps. Some other works used adaptive optimization of the number of informative tokens to improve efficiency [Wang *et al.*, 2021b] [Rao *et al.*, 2021] [Yin *et al.*, 2022] [Xu *et al.*, 2022]. Zhang et al. [Zhang *et al.*, 2022a] [Zhang *et al.*, 2022b] designed their Trans4PASS and Trans4PASS+ based on DPE and Deformable Multi-Layer Perception (DMLP) modules, and we use Trans4PASS+ as our baseline.

3 Method

We present Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation (SGAT4PASS) in this section. First, we introduce the background of panoramic semantic segmentation in Section 3.1. Second, we describe our main idea to apply different SGA properties in panoramic semantic segmentation in Section 3.2. To improve the 3D robustness of SGAT4PASS, we propose SGA Image Projection, Spherical Deformable Patch Embedding (SDPE), and panorama-aware loss. Specifically, SGA Image Projection adds rotated samples in training; SDPE adds SGA constraints on the deformable patch embedding; and the panorama-aware loss fuses sphere pixel density to training process.

3.1 Background

We first describe a general formulation of PASS and then introduce the spherical geometry property that we focus mainly on. Panoramic images are based on original 360° data formulated in the spherical coordinate system (based on longitude and latitude). To convert it to a rectangular image in a Cartesian coordinate system, ERP is a widely used projection in this field: $x = (\theta - \theta_0)cos\phi_1, y = (\phi - \phi_1)$, where $\theta_0 = 0$ is the central latitude and $\phi_1 = 0$ is the central longitude. The ERP-processed rectangular images are used as the input sample in datasets and fed to the neural network, and the rectangular semantic segmentation results are obtained to compare with the ground truth and calculate the metrics. Although traditional methods can treat PASS as the conventional 2D semantic segmentation task and deal with panoramic images easily, the spherical geometry property is partly ignored.

3.2 Spherical Geometry-Aware (SGA) Framework

We propose the SGA framework for PASS with SGA image projection, SDPE, and panorama-aware loss. To deal with the inevitable 3D disturbance during the acquisition of the input image, our SGA image projection aims to encode the original 360° data spherical geometry by generating input images with different rotations. We design SDPE to model spatial dependencies on a sphere, making patch embedding consider

Method	Avg mIoU	F1 mIoU
StdConv [Tateno *et al.*, 2018]	-	32.6
CubeMap [Tateno *et al.*, 2018]	-	33.8
DistConv [Tateno *et al.*, 2018]	-	34.6
SWSCNN [Esteves *et al.*, 2020]	43.4	-
Tangent (ResNet-101) [Eder *et al.*, 2020]	45.6	-
FreDSNet [Berenguel-Baeta *et al.*, 2022]	-	46.1
PanoFormer [Shen *et al.*, 2022]	48.9	-
HoHoNet (ResNet-101) [Sun *et al.*, 2021]	52.0	53.9
Trans4PASS (Small) [Zhang *et al.*, 2022a]	52.1	53.3
CBFC [Zheng *et al.*, 2023]	52.2	-
Trans4PASS+ (Small) [Zhang *et al.*, 2022b]	53.7	53.6
Ours (Small)	**55.3**	**56.4**

Table 1: Comparison with the SOTA methods on Stanford2D3D Panoramic datasets. We follow recent works to compare the performance of both official fold 1 and the average performance of all three official folds. respectively. "Avg mIoU" / "F1 mIoU" means the mIoU performance of three official folds on average / official fold 1. A considerable improvement is gained.

both spherical geometry and image distortions. Furthermore, a panorama-aware loss is proposed to model the pixel density of a sphere, making the loss weight distribution more similar to the original 360° data. With these three modules, the spherical geometry is well employed in the PASS task.

Spherical Geometry-Aware (SGA) Image Projection

The original 360° data follow a spherical distribution and are *spherically symmetric*. After rotating any angle along the yaw / pitch / roll axis, the transformed data are still equivalent to the original data. Traditional strategies assume that the images are taken with the yaw / pitch / roll angle equal to zero degrees, which is too ideal in real-world scenarios and ignores the camera disturbance and random noise. When the rotation angle is disturbed, traditional strategies usually have a large degradation in the PASS task. SGA image projection fuses this property between the inevitable equirectangular projection and regular image augmentation to make models robust to 3D disturbance.

We use T to represent the forward process of ERP transformation, which is the process of converting spherical coordinates to plane coordinates, and use T^{-1} to represent the inverse process of ERP that transforms the plane back onto the sphere. Given an ERP-processed input panoramic image, we first transform the image I originally in plane coordinates to spherical coordinates through the inverse ERP process. After that, we use the rotation matrix in the three-dimensional (3D) space to perform a 3D rotation in the spherical coordinate system. For a general rotation in a 3D space, the angles of yaw, pitch, and roll are α_{use}, β_{use}, and γ_{use}, respectively. The corresponding rotation matrix is $R(\alpha_{\text{use}}, \beta_{\text{use}}, \gamma_{\text{use}})$. We multiply R by the data in the spherical coordinate system to obtain the rotated data in the spherical coordinate system. Finally, we use the ERP forward process to convert the rotated spherical coordinate system image into a panoramic image, thus obtaining a certain rotated image of the real input of the network. The corresponding point in input image of a pixel in rotated image may not have integer coordinates, and we select the nearest pixel as its corresponding pixel to be generic to the ground truth transformation. Based on these operations, we build our SGA image projection, $O_{3D}(I, \alpha_{\text{use}}, \beta_{\text{use}}, \gamma_{\text{use}}) = T(R(\alpha_{\text{use}}, \beta_{\text{use}}, \gamma_{\text{use}}) \cdot T^{-1}(I))$. (See Section C "Details for SGA Image Projection" in the supplementary material for details.) At the beginning of the training process, we set the maximum rotation angle of the yaw / pitch / roll axis at $(\alpha_{\text{train}}, \beta_{\text{train}}, \gamma_{\text{train}})$.

SDPE: Spherical Deformable Patch Embedding

We first introduce DPE, and then fuse spherical geometry into DPE by SGA constraints to earn SDPE.

Faced with image distortions in panoramic images, DPE, considering different distortions in different regions of an input panoramic image, is a popular solution [Zhang *et al.*, 2022a] [Zhang *et al.*, 2022b]. In detail, given a 2D input panoramic image, the standard patch embedding handles it into flattened patches $H \times W$, and the resolution of each patch is (s, s). A learnable projection layer transforms each patch into out-dimensional embeddings. For each patch, the offsets $\Delta^{DPE}_{(i,j)}$ of the i^{th} row j^{th} column pixel are defined as:

$$\Delta^{DPE}_{(i,j)} = \begin{bmatrix} \min(\max(\text{-}k_D \cdot H, g(f)_{(i,j)}), k_D \cdot H) \\ \min(\max(\text{-}k_D \cdot W, g(f)_{(i,j)}), k_D \cdot W) \end{bmatrix}, \quad (1)$$

where $g(\cdot)$ is the offset prediction function. Hyperparameter k_D puts an upper bound on the learnable offsets $\Delta^{DPE}_{(i,j)}$. For implementation, the deformable convolution operation [Dai *et al.*, 2017] is popularly employed to realize DPE.

When fusing spherical geometry into DPE, human photographic and ERP priors are taken into consideration, in which the plane formed by pitch and roll axes is always parallel to the ground plane and the projection cylinder is perpendicular to the ground plane. As a result, we add SGA constraints mainly on the yaw axis. In detail, we give intra-offset and inter-offset constraints on $\Delta^{DPE}_{(i,j)}$. For convenience, we use $\Delta^{m,n}_{(i,j)}$ to represent the i^{th} row j^{th} column pixel of the learnable offset for the m^{th} row n^{th} column patch.

Intra-offset constraint. Based on the phenomenon that the original 360° data are symmetric on any longitude and the projection cylinder in ERP is symmetric in any line perpendicular to the base of the cylinder, the offset of any pixel in 2D input panoramic image I should be symmetric on its perpendicular. To be generic to the learnable offsets $\Delta^{m,n}_{(i,j)}$ dealing with the image distortions, we use a constraint $\mathcal{L}_{intra}$:

$$\mathcal{L}_{intra} = \sum_{m,n} \sum_{i,j} L_2^{intra}(\Delta^{m,n}_{(i,j)}, \Delta_S{}^{m,n}_{(i,j)}), \quad (2)$$

where $\Delta_S{}^{m,n}_{(i,j)}$ is the single patch offset that is formed symmetrically along the yaw axis with $\Delta^{m,n}_{(i,j)}$ as the template. $L_2^{intra}(\cdot,\cdot)$ represents the element-wise L2 loss.

Inter-offset constraint. Based on the phenomenon that the projection cylinder in ERP can be slit and expanded from any line perpendicular to the base of the cylinder, the offset of any pixel in 2D input panoramic image I corresponding to the same latitude of the original 360° data should be similar. To be generic to the learnable $\Delta^{DPE}_{(i,j)}$ dealing with the image distortions, we use a constraint, $\mathcal{L}_{inter}$, to model this property. For a certain pixel, we use the average offset in the

(β,γ,α) (°)	BL mIoU / PAcc Our mIoU / PAcc	(β,γ,α) (°)	BL mIoU / PAcc Our mIoU / PAcc	(β,γ,α) (°)	BL mIoU / PAcc Our mIoU / PAcc	(β,γ,α) (°)	BL mIoU / PAcc Our mIoU / PAcc
(0,0,0)	53.617 / 81.483 56.374 / 83.135	(0,5,0)	49.292 / 78.346 56.073 / 82.892	(5,0,0)	49.468 / 78.500 56.074 / 82.905	(5,5,0)	47.234 / 77.129 55.784 / 82.794
(0,0,90)	53.918 / 81.590 56.441 / 83.130	(0,5,90)	49.861 / 78.656 55.954 / 82.847	(5,0,90)	49.400 / 78.373 56.128 / 82.895	(5,5,90)	47.589 / 77.361 55.636 / 82.657
(0,0,180)	53.587 / 81.476 56.246 / 83.054	(0,5,180)	49.344 / 78.532 55.951 / 82.906	(5,0,180)	49.536 / 78.585 55.714 / 82.796	(5,5,180)	47.458 / 77.307 55.501 / 82.750
(0,0,270)	53.669 / 81.459 56.223 / 83.051	(0,5,270)	49.462 / 78.445 55.924 / 82.779	(5,0,270)	49.363 / 78.485 55.983 / 82.904	(5,5,270)	47.726 / 77.451 55.732 / 82.701

Table 2: Detail performance comparison with Tran4PASS+ on Stanford2D3D Panoramic datasets official fold 1 with SGA metrics. All 18 situations are shown, and the analysis is in table 3. "BL" means the baseline, i.e., Tran4PASS+. "PAcc" meas the pixel accuracy metric.

Statistics	Baseline		Ours	
	mIoU	PAcc	mIoU	PAcc
Mean	50.033	78.949	55.984 (+5.951)	82.887 (+3.938)
Variance	5.147	2.413	0.066 (-5.081)	0.020 (-2.393)
Range	6.684	4.461	0.940 (-5.744)	0.478 (-3.983)

Table 3: Overall performance comparison with Tran4PASS+ on Stanford2D3D Panoramic datasets in table 2 setting. "PAcc" means the pixel accuracy metric. SGAT4PASS earns considerable mean performance and significant robustness improvement.

whole horizontal line as its constraint:

$$\mathcal{L}_{inter} = \sum_{m,n} \sum_{i,j} L_2^{inter}(\Delta_{(i,j)}^{m,n}, \Delta_{(i,j)}^{m,\text{AVG}}), \tag{3}$$

where $\Delta_{(i,j)}^{m,\text{AVG}}$ is the average of each component in $\{\Delta_{(i,j)}^{m,n}, n \in W\}$, and $L_2^{inter}(\cdot,\cdot)$ represents the L2 loss for each component length of the two vectors. Then the total SDPE loss is: $\mathcal{L}_{SDPE} = \mathcal{L}_{inter} + \mathcal{L}_{intra}$.

Panorama-Aware Loss

Because the panoramic images are rectangular in shape, the region of the antarctic and arctic areas in the original 360° data is over sampled than the one near the equator. However, due to human photographic priors, the semantics of the antarctic (ground, floor, etc.) and arctic areas (sky, ceiling, etc.) are relatively simple, as seen in the sample images of Figure 1 and Figure 2. When using traditional segmentation loss for supervised training, we treat each pixel equally, which leads to models paying relatively less attention to semantic rich regions near the equator. To deal with this phenomenon, we design our panorama-aware loss. For an ERP-processed panoramic image, the number of pixels in each horizontal line is the same, but the corresponding resolution density on the original sphere of each horizontal line is very different. For this reason, we design a loss to reweight the loss proportion of different horizontal lines depending on its height. For a pixel $(m,n)|m \in [1,H_I], n \in [1,W_I]$ (W_I and H_I are the width and height of the input image), we give a weight $w_{\text{pan}}^{(m,n)}$ when calculating its per pixel loss:

$$w_{\text{pan}}^{(m,n)} = \cos(\frac{|2m - H_I|}{H_I} \cdot \frac{\pi}{2}). \tag{4}$$

We use W_{pan} to represent the set that includes all $w_{\text{pan}}^{(m,n)}$.

When faced with a panoramic semantic segmentation problem, we first estimate the usage scenario to determine β and γ used in SGA image projection when α is often set as 360° in common condition. We set our total loss as:

$$\mathcal{L}_{\text{all}} = (1 + \lambda_w \cdot W_{\text{pan}}) \odot \mathcal{L}_{SEG} + \lambda_s \cdot \mathcal{L}_{SDPE}, \tag{5}$$

where $\mathcal{L}_{SEG}$ is the common per pixel loss for semantic segmentation, $\odot$ is the element-wise matrix multiplication, λ_w and λ_s are hyperparameters.

SGAIP	SDPE	PA	mIoU	Pixel accuracy
			53.617	81.483
✓			54.637	82.303
	✓		54.554	81.508
		✓	54.833	81.733
✓	✓	✓	56.374	83.135

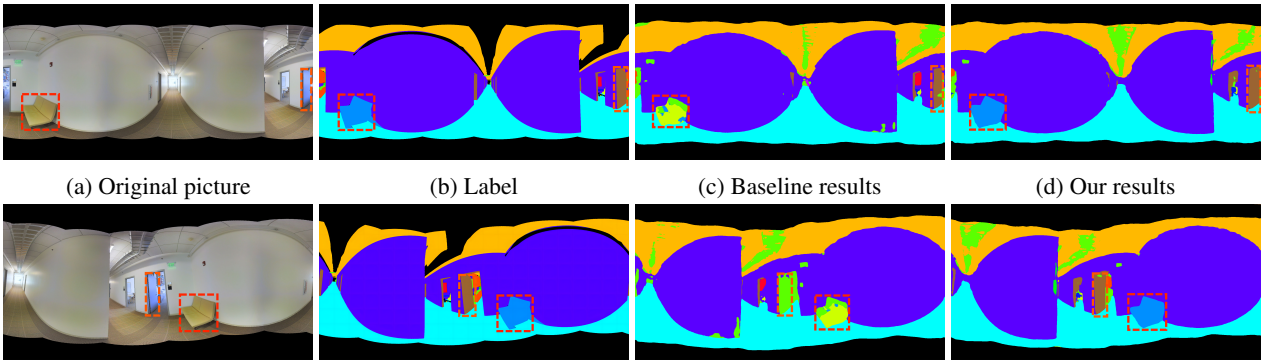
Table 4: Effect of each SGAT4PASS module. We validate them on Stanford2D3D Panoramic datasets official fold 1 with traditional metrics. "SGAIP" / "SDPE" / "PA" means our SGA image projection / spherical deformable patch embedding / panorama-aware loss. Using anyone, an average improvement of 1.058% mIoU / 0.365% pixel accuracy is gained when using three gains 2.757% / 1.652%.

4 Experiments

In this section, we evaluate our SGAT4PASS against the popular benchmark, Stanford2D3D, for both traditional metrics and our SGA validation.

4.1 Datasets and Protocols

We validate SGAT4PASS on Stanford2D3D Panoramic datasets [Armeni *et al.*, 2017]. It has 1,413 panoramas, and



(a) Original picture (b) Label (c) Baseline results (d) Our results

(e) Rotated original picture (f) Rotated label (g) Baseline rotated results (h) Our rotated results

Figure 3: Visualization comparison of SGAT4PASS and Trans4PASS+. The rotation of the pitch / roll / yaw axis is 5° / 5° / 180°. SGAT4PASS gains the better results of semantic class "door" and "sofa" (highlighted by red dotted line boxes).

Statistics	Baseline		Ours	
	mIoU	Pixel accuracy	mIoU	Pixel accuracy
$(\beta, \gamma, \alpha) = (1°, 1°, 360°)$				
Mean	53.473	81.251	56.212 (+2.739)	83.021 (+1.770)
Variance	0.056	0.029	0.011 (-0.045)	0.003 (-0.026)
Range	0.856	0.591	0.394 (-0.462)	0.192 (-0.399)
$(\beta, \gamma, \alpha) = (0°, 0°, 360°)$				
Mean	53.698	81.502	56.321 (+2.623)	83.093 (+1.591)
Variance	0.017	0.003	0.008 (-0.009)	0.002 (-0.001)
Range	0.331	0.131	0.218 (-0.113)	0.084 (-0.047)

Table 5: Overall performance comparison on Stanford2D3D Panoramic datasets in different SGA metrics in two more favorable settings for Tran4PASS+. SGAT4PASS also earns considerable mean performance and significant robustness improvement.

13 semantic classes are labeled, and has 3 official folds, fold 1 / 2 / 3. We follow the report style of previous work [Zhang *et al.*, 2022a] [Zhang *et al.*, 2022b].

Our experiments are conducted with a server with four A100 GPUs. We use Trans4PASS+ [Zhang *et al.*, 2022b] as our baseline and set an initial learning rate of 8e-5, which is scheduled by the poly strategy with 0.9 power over 150 epochs. The optimizer is AdamW [Kingma and Ba, 2015] with epsilon 1e-8, weight decay 1e-4, and batch size is 4 on each GPU. Other settings and hyperparameters are set the same as Trans4PASS+ [Zhang *et al.*, 2022b]. For each input panoramic image I in an iteration, there is a 50% chance of using it directly and the other 50% chance of using it after SGA image projection, $O_{3D}(I, \alpha_{\text{use}}, \beta_{\text{use}}, \gamma_{\text{use}})$, where α_{use} / β_{use} / γ_{use} uniformly sampled from 0 to α_{train} / β_{train} / γ_{train}. We set $(\beta_{\text{train}}, \gamma_{\text{train}}, \alpha_{\text{train}}) = (10°, 10°, 360°)$. λ_w and λ_s are set as 0.3 and 0.3, respectively.

SGA validation. Most PASS datasets use a unified ERP way to process original 360° data, PASS models have the potential to overfit the ERP way, cannot handle 3D disturbance well and have little 3D robustness. To validate the robustness of the PASS models, we propose a novel SGA validation. n_α, n_β, and n_γ are the number of different angles for the yaw / pitch / roll axis, respectively, and $n_\alpha \cdot n_\beta \cdot n_\gamma$ different-angle panoramic images for a certain original 360° data is earned. Panoramic semantic segmentation models are validated in all $n_\alpha \cdot n_\beta \cdot n_\gamma$ settings, and their statistics are reported as SGA metrics. In our SGA validation, "Mean" means the average of all $n_\alpha \cdot n_\beta \cdot n_\gamma$ traditional results (e.g., mIoU, per pixel accuracy, etc.). "Variance" means the variance of all $n_\alpha \cdot n_\beta \cdot n_\gamma$ results. "Range" means the gap between the maximum and minimum results of all $n_\alpha \cdot n_\beta \cdot n_\gamma$ results. Compared to traditional validation, SGA validation avoids models gain performance by fitting the ERP way of datasets and reflects objective 3D robustness. In detail, we assume that the 3D rotation disturbance is at most 5° / 5° / 360° of pitch (β) / roll (γ) / yaw (α) angle. We set $n_\alpha = 4$ $(0°, 90°, 180°, 270°)$, $n_\beta = 2$ $(0°, 5°)$, and $n_\gamma = 2$ $(0°, 5°)$. We use the mean of them as the final performance and observe the performance difference among them to indicate the 3D robustness of models.

4.2 Performance Comparison

In this part, we first compare several recent SOTA methods with traditional metrics, and then compare the latest SOTA Trans4PASS+ in detail with SGA metrics.

Traditional metrics. Comparison results on Stanford2D3D Panoramic datasets with SOTA methods in traditional metrics are shown in Table 1. Following recent work, we report the performance of both official fold 1 and the average performance of all three official folds. From the results, SGAT4PASS outperforms current SOTA models by 2.8% / 1.6% mIoU, respectively, which means that our SGAT4PASS has a considerable performance margin compared to current models with traditional metrics.

SGA metrics. Comparison results on Stanford2D3D Panoramic datasets with our SGA validation metrics are shown in Table 3, and Table 2 is the detailed performance of each situation. For mean mIoU / pixel accuracy, an improvement of nearly 6% / 4% is achieved, respectively. Furthermore, our variance is about $\frac{1}{100}$ and our fluctuation range is about $\frac{1}{10}$. These results show that our SGAT4PASS have much better robustness than Trans4PASS+.

Network	Test Method	mIoU	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Trans4Pass+	Traditional	53.62	0.39	74.4	65.32	84.21	62.86	36.44	15.96	32.79	93.09	44.10	63.67	75.02	46.90
Ours		56.37	0.73	74.05	65.91	84.20	64.53	41.24	19.62	**52.67**	93.08	**56.92**	58.86	76.43	44.62
Trans4Pass+	SGA	50.03	0.26	73.78	62.21	83.82	61.87	32.11	10.93	20.26	92.96	38.33	61.78	74.35	37.73
Ours		55.98	0.78	73.94	65.56	84.08	64.39	40.96	18.31	**51.64**	92.98	**56.53**	58.14	76.06	44.42

Table 6: Per-class mIoU results on Stanford2D3D Panoramic datasets according to the fold 1 data setting with traditional mIoU and per-pixel accuracy metrics. No mark for the results that the gap between Trans4Pass+ and Ours less than 5% (performance at the same level). Our results will be red when Ours outperforms more than 5%. If Ours outperforms more than 10%, our results will be **bold and red**. There is no semantic class that Trans4Pass+ outperforms Ours 5% or more.

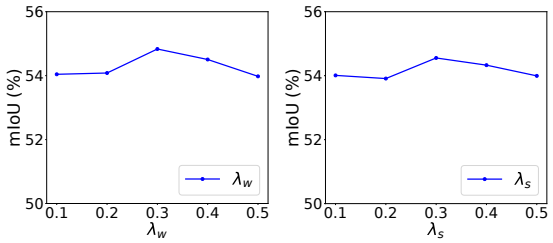
4.3 Ablation Study

Effect of three modules in training process. The effectiveness of SGA image projection, SDPE, and panorama-aware loss are studied on Stanford2D3D Panoramic datasets official fold 1 with traditional metrics as shown in Table 4. (a) SGA image projection: Using it alone improves the baseline mIoU / per pixel accuracy by 1.020% / 0.820%. (b) SDPE: Using SDPE alone outperforms the baseline by 0.937% and 0.025% in mIoU and per pixel accuracy. (b) Panorama-aware loss: Using it alone improves the baseline by 1.216% and 0.250% in mIoU and per pixel accuracy.

Effect of SGA validation. We demonstrate the effect of SGA validation, which means a stronger generalizability to resist 3D rotational perturbation. We carried out experiments with two smaller disturbance settings on the pitch and roll axes ($(\beta, \gamma, \alpha) = (1°, 1°, 360°)$ / $(0°, 0°, 360°)$), which are more favorable settings for Trans4PASS+ [Zhang *et al.*, 2022b], because it is designed for the standard panoramic view image ($(\beta_{\text{use}}, \gamma_{\text{use}}, \alpha_{\text{use}}) = (0°, 0°, 0°)$). The overall statistical results are shown in Table 5. For the $(\beta, \gamma, \alpha) = (1°, 1°, 360°)$ setting, an improvement of approximately 2.7 % / 1.7 % is obtained for the mean mIoU / pixel accuracy. Our variance is approximately $\frac{1}{5}$ / $\frac{1}{10}$ and our fluctuation range is approximately $\frac{1}{2}$ / $\frac{1}{3}$ in mIoU / pixel precision. In $(\beta, \gamma, \alpha) = (0°, 0°, 360°)$ setting, mean mIoU / pixel accuracy gains approximately 2.6% / 1.6% improvement, variances / fluctuation is approximately $\frac{1}{2}$ / $\frac{2}{3}$ for SGAT4PASS. SGAT4PASS has better robustness even with little 3D perturbations. The detailed performance of these two settings and the performance of several random rotation settings are shown in Section A "Detailed Performance of SGA Validation" in the supplementary material.

4.4 Discussion and Visualizations

Performance of all semantic classes and visualizations. We show the detailed performance of all 13 semantic classes on the Stanford2D3D Panoramic datasets with both traditional and SGA metrics in Table 6, respectively. We focus mainly on the classes with significant performance gaps and mark the gap larger than 5% / 10% as red numbers / bold red numbers, respectively. There is no semantic class for which the baseline is significantly better. From the results, we can learn that the "sofa" and "door" classes improve more. An image with "door" and "sofa" is visualized in Figure 3. Rotation of the pitch / roll / yaw axis is 5° / 5° / 180°. The baseline prediction gap between the original and rotated input is large, which means less robustness. It predicts the door near the right boundary in Figure 3c overall right, but it is totally wrong with rotation in Figure 3g when SGAT4PASS predicts both correct. The baseline predictions for the sofa change a lot with rotation when SGAT4PASS is stable. More visualizations are shown in Section B "More Visualizations" in the supplementary material.



(a) λ_w-mIoU (b) λ_s-mIoU

Figure 4: Influence of λ_s and λ_w in SGAT4PASS. The results are carried out on Stanford2D3D Panoramic datasets official fold 1.

Different hyperparameters. λ_w and λ_s are hyperparameters in our SGAT4PASS. λ_s / λ_w determines the proportion of our constraint of spherical geometry in SDPE / panorama-aware loss. We apply them on the baseline, respectively. Traditional mIoU results are shown in Figure 4a and Figure 4b, and we choose 0.3 / 0.3 as the final λ_w / λ_s.

5 Conclusion

We have studied an underexplored but important field in panoramic semantic segmentation, i.e., the robustness of dealing with 3D disturbance panoramic input images. We have shown that using our SGA framework is key to improving the semantic segmentation quality of 3D disturbance inputs. It applies spherical geometry prior to panoramic semantic segmentation and gains considerable improvement. In detail, the SGA framework includes SGA image projection, SDPE, and panorama-aware loss. We also validated the effectiveness of our SGAT4PASS on popular datasets with the traditional metrics and the proposed SGA metrics, and studied its properties both empirically and theoretically.

Acknowledgements

This work is supported in part by National Key Research and Development Program of China under Grant 2020AAA0107400, National Natural Science Foundation of China under Grant U20A20222, National Science Foundation for Distinguished Young Scholars under Grant 62225605, Research Fund of ARC Lab, Tencent PCG, The Ng Teng Fong Charitable Foundation in the form of ZJU-SUTD IDEA Grant, 188170-11102 as well as CCF-Zhipu AI Large Model Fund (CCF-Zhipu202302).

Contribution Statement

Xuewei Li and Tao Wu contributed equally to this work.

References

[Ai *et al.*, 2022] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Ling Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022.

[Armeni *et al.*, 2017] 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.

[Berenguel-Baeta *et al.*, 2022] Bruno Berenguel-Baeta, Jesus Bermudez-Cameo, and Jose J Guerrero. Fredsnet: Joint monocular depth and semantic segmentation with fast fourier convolutions. *arXiv preprint arXiv:2210.01595*, 2022.

[Chen *et al.*, 2021] Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. Dpt: Deformable patch-based transformer for visual recognition. In *Proc. ACM MM*, 2021.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. ICCV*, pages 764–773, 2017.

[de La Garanderie *et al.*, 2018] Greire 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 *Proc. ECCV*, pages 789–807, 2018.

[Deng *et al.*, 2017] Liuyuan Deng, Ming Yang, Yeqiang Qian, Chunxiang Wang, and Bing Wang. Cnn based semantic segmentation for urban traffic scenes using fisheye camera. In *Proc. IEEE Intell. Vehicles Symp.*, pages 231–236. IEEE, 2017.

[Eder *et al.*, 2020] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proc. CVPR*, 2020.

[Esteves *et al.*, 2020] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical cnns. *Proc. NeurIPS*, 33:8614–8625, 2020.

[Gao *et al.*, 2022] Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. Review on panoramic imaging and its applications in scene understanding. *arXiv preprint arXiv:2205.05570*, 2022.

[Hu *et al.*, 2022] Xing Hu, Yi An, Cheng Shao, and Huosheng Hu. Distortion convolution module for semantic segmentation of panoramic images based on the image-forming principle. *IEEE Trans. Instrum. Meas.*, 71:1–12, 2022.

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

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

[Lee *et al.*, 2018] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360 degree images. *arXiv preprint arXiv:1811.08196*, 2018.

[Ma *et al.*, 2021] 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 *Proc. ITSC*, pages 2766–2772. IEEE, 2021.

[Pintore *et al.*, 2021] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR*, pages 11536–11545, 2021.

[Poulin-Girard and Thibault, 2012] Anne-Sophie Poulin-Girard and Simon Thibault. Optical testing of panoramic lenses. *Opt. Eng.*, 51(5):053603, 2012.

[Rao *et al.*, 2021] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Proc. NeurIPS*, 34:13937–13949, 2021.

[Shen *et al.*, 2022] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *Proc. ECCV*, pages 195–211. Springer, 2022.

[Sun *et al.*, 2021] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proc. CVPR*, pages 2573–2582, 2021.

[Tateno *et al.*, 2018] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proc. ECCV*, pages 707–722, 2018.

[Wang *et al.*, 2021a] 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 *Proc. ICCV*, pages 568–578, 2021.

[Wang *et al.*, 2021b] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth

16x16 words: Dynamic transformers for efficient image recognition. *Proc. NeurIPS*, 34:11960–11973, 2021.

[Xia *et al.*, 2022] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proc. CVPR*, pages 4794–4803, 2022.

[Xu *et al.*, 2018] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2693–2708, 2018.

[Xu *et al.*, 2021] Yanyu Xu, Ziheng Zhang, and Shenghua Gao. Spherical dnns and their applications in 360 images and videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[Xu *et al.*, 2022] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proc. AAAI*, volume 36, pages 2964–2972, 2022.

[Yang *et al.*, 2019] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, and Kaiwei Wang. Pass: Panoramic annular semantic segmentation. *IEEE Trans. Intell. Trans. Syst.*, 21(10):4171–4185, 2019.

[Yang *et al.*, 2020] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelhagen. Dspass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. In *Proc. IEEE Intell. Vehicles Symp.*, pages 457–464. IEEE, 2020.

[Yang *et al.*, 2021a] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Trans. Image Process.*, 30:1866–1881, 2021.

[Yang *et al.*, 2021b] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proc. CVPR*, pages 1376–1386, 2021.

[Ye *et al.*, 2021] Zhiyuan Ye, Hai-Bo Wang, Jun Xiong, and Kaige Wang. Ghost panorama using a convex mirror. *Opt. Lett.*, 46(21):5389–5392, 2021.

[Yin *et al.*, 2022] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proc. CVPR*, pages 10809–10818, 2022.

[Yue *et al.*, 2021] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proc. ICCV*, pages 387–396, 2021.

[Zhang *et al.*, 2022a] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proc. CVPR*, pages 16917–16927, 2022.

[Zhang *et al.*, 2022b] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022.

[Zheng *et al.*, 2023] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360deg semantic segmentation with self-distillation. In *Proc. WACV*, pages 4501–4510, 2023.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[Zhuang *et al.*, 2022] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *Proc. AAAI*, volume 36, pages 3653–3661, 2022.