Supplementary Materials for Open Panoramic Segmentation

Junwei Zheng[1], Ruiping Liu[1], Yufan Chen[1], Kunyu Peng[1], Chengzhi Wu[1], Kailun Yang[2], Jiaming Zhang[1,†], and Rainer Stiefelhagen[1]

[1] Karlsruhe Institute of Technology
[2] Hunan University

A Implementation Details

Hardware Setup. In this work, we train our models using 4×A40 GPUs with respective 40 GB memory. This computational node also contains 200 GB CPU memory. The source code and the model checkpoints will be made publicly available.

Training Settings. The training process utilizes the AdamW optimizer, employing an initial learning rate of 0.0001, a weight decay of 0.0001, a batch size of 32, and a total of 60,000 training iterations. During the training phase, the learning rate adheres to a polynomial schedule with a power of 0.9. Apart from the training specification listed in Table 1a, Table 1b illustrates the data augmentations and the corresponding parameters used in the training. The short edge of an image is randomly resized in a range of [320, 1024]. Afterward, the image is randomly cropped into 640×640. For RandomBrightness, RandomContrast, RandomSaturation, and RandomHue, they are randomly applied with a probability of 0.5. The brightness delta, contrast range, saturation range, and hue delta of the aforementioned data augmentations are reported in Table 1b. The input image is also flipped in the horizontal direction with 0.5 probability before being forwarded to the model. The whole training process takes roughly 10 hours.

Table 1: Implementation details.

(a) Training settings.

Configurations	Parameter
Optimizer	AdamW
Learning Rate	0.0001
Weight Decay	0.0001
Scheduler	Poly. (power 0.9)
Training Iterations	60,000
Batch Size per GPU	8

(b) Data augmentation during the training process.

Configurations	Parameter
RandomResize	[320, 1024]
RandomCrop	640 × 640
RandomBrightness	32
RandomContrast	[0.5, 1.5]
RandomSaturation	[0.5, 1.5]
RandomHue	18
RandomFlip	Horizontal

† Correspondence: jiaming.zhang@kit.edu

B Qualitative Results

B.1 Visualization of Segmentation Predictions

Fig. 1 presents the visualization of segmentation predictions on WildPASS [3] dataset. The first row is the input panoramic images. The second row is the close-vocabulary segmentation predictions with only 8 predefined categories in the WildPASS dataset while the last row is the open-vocabulary segmentation predictions with an arbitrary number of classes. It can be observed that all pixels of the entire image have their own semantic meanings, showcasing the superiority of our proposed OOOPS model and the zero-shot learning ability. It is worth noting that even the challenging category, *e.g.*, mailbox, can be detected in the open-vocabulary setting. Fig. 2 and Fig. 3 illustrate the visualization of segmentation predictions on Stanford2D3D and Matterport3D datasets, respectively. It's obvious that all predefined categories of these two datasets can be predicted correctly by the OOOPS model. Beyond the correctness, the object deformations, *e.g.*, the door in the middle column of Fig. 2, can also be detected by our proposed OOOPS model, indicating the OOOPS model is aware of the image distortion and object deformation.

B.2 Visualization of Deformable Offsets

Since our OOOPS model is specifically designed for image distortion and object deformation, it is necessary to present the deformation-aware capability of the model. Fig. 4, Fig. 5 and Fig. 6 illustrate the deformable offsets on WilPASS, Stanford2D3D and Matterport3D dataset, respectively. The green points • are the sample locations. The red points • are deformable offsets in 2 levels, indicating a deformable receptive field (*e.g.*, each level has a 3×3 kernel size, resulting in $(3 \times 3)^2 = 81$ red points). Leveraging the standard deviation of the cosine similarity vector calculated by the center pixel and all pixels within a kernel, the OOOPS model is capable of capturing the salient pixels, *e.g.*, edge pixels of an image where strong panoramic distortion usually occurs. For example, the sidewalk in the first row and second column of Fig. 4 has a very strong distortion due to the Equirectangular Projection [2] (ERP) from a globe to a plane resulting in a panoramic image. Although the image distortion occurs, the green sample location has a deformable receptive field presented by red points along the edges of the sidewalk, indicating the deformation-aware capability of the OOOPS model. The deformable awareness can be observed not only in the outdoor WildPASS panoramic dataset in Fig. 4 but also in the indoor Stanford2D3D and Matterport3D datasets in Fig. 5 and Fig. 6.

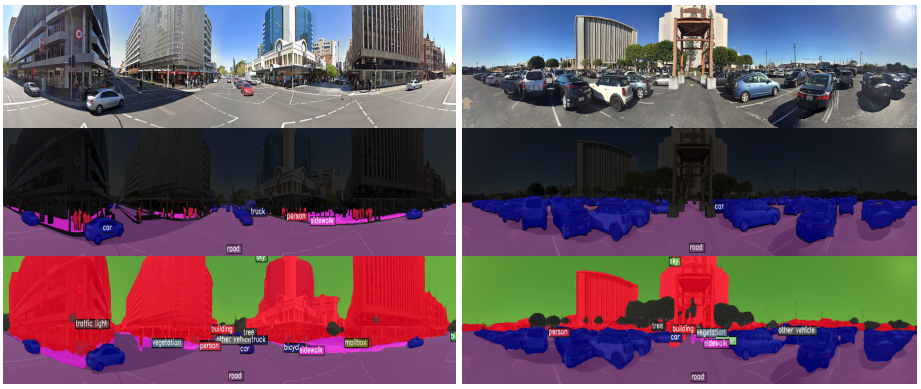

Fig. 1: Visualization on the WildPASS dataset. First row: RGB images. Second row: close-vocabulary predictions of the proposed OOOPS model. Third row: open-vocabulary predictions of the proposed OOOPS model.

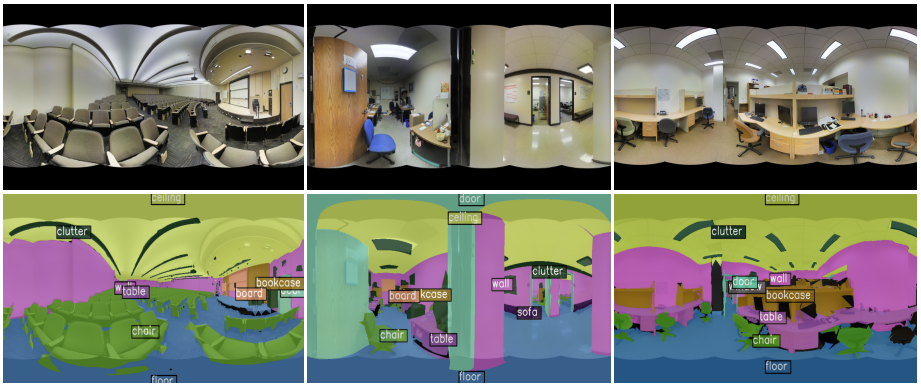

Fig. 2: Visualization on the Stanford2D3D dataset. First row: RGB images. Second row: predictions of the proposed OOOPS model.

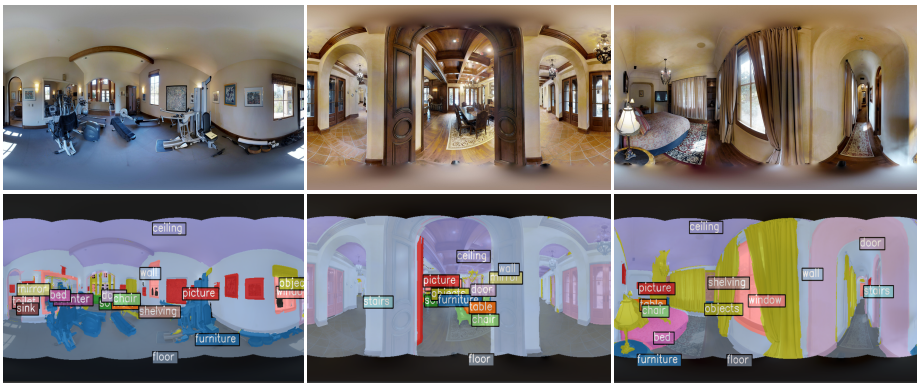

Fig. 3: Visualization on the Matterport3D dataset. First row: RGB images. Second row: predictions of the proposed OOOPS model.

Fig. 4: Visualization of deformable offsets on WildPASS dataset. The green points • are the sample locations. The red points • are deformable offsets in 2 levels, indicating a deformable receptive field (*e.g.*, each level has a 3×3 kernel size, resulting in $(3 \times 3)^2 = 81$ red points). Zoom in for a better view.

Fig. 5: Visualization of deformable offsets on Stanford2D3D dataset.

Fig. 6: Visualization of deformable offsets on Matterport3D dataset.

Table 2: mIoU of RERP with adaptive sizes on WildPASS, Stanford2D3D, and Matterport3D datasets. mIoU is in percentage (%).

Method	WildPASS	Stanford2D3D	Matterport3D
RERP	58.0	41.1	31.2
RERP w/ adaptive sizes	**58.5**	**41.5**	**31.6**

Table 3: mIoU of RERP and simple horizontal rotation on WildPASS, Stanford2D3D, and Matterport3D datasets. mIoU is in percentage (%).

Method	WildPASS	Stanford2D3D	Matterport3D
OOOPS w/o RERP	57.0	39.5	31.1
OOOPS w/ Rotation	57.0	39.5	31.1
OOOPS w/ RERP	**58.0**	**41.1**	**31.2**

C Ablation Study

C.1 Adaptive Shuffling Patches

We conduct an experiment that divides pinhole images into parts of adaptive sizes when doing RERP augmentation, similar to Mosaic augmentation. The results in Table 2 indicate adaptive sizes can further boost the model performance.

C.2 Simple Horizontal Rotation

The simple horizontal rotation [1] is used to get a new panorama with a different viewpoint. It is applied to panoramas, not pinhole images. RERP is used to transform a pinhole image into a panorama-like image. It is applied to pinhole images, not panoramas. We experiment with applying simple horizontal rotation to pinhole images. From Table 3 we find that simple horizontal rotation does not bring gains, which falls behind the one with RERP.

D Discussion

Limitations and Future Work. In this work, we focus on the open panoramic segmentation, where the models are trained in the narrow-FoV pinhole source domain in an open-vocabulary setting while evaluated in the wide-FoV panoramic target domain. Compared to the state-of-the-art methods trained in a close-vocabulary setting, the limitations of the proposed OOOPS model are obvious. The performance of the open-vocabulary model falls short of the close-vocabulary ones. The architectural design does not encompass the 360° boundaries of panoramas, providing an opportunity for improving seamless scene segmentation. Additionally, the generalization capability of the OOOPS model can

be evaluated using surround-view fisheye images. Our future plans involve extending the proposed solution to encompass panoramic panoptic segmentation.
Societal Impacts. The proposed Open Panoramic Segmentation (OPS) task and the OOOPS model with Random Equirectangular Projection (RERP) enable distortion-aware open-vocabulary panoramic semantic segmentation in different open domains even though there are no training-sufficient dense-annotated panoramic labels. Evidently, this represents a great technological advancement that necessitates not only strategic utilization but also a thorough awareness of the inherent risks. Although this work is able to predict an arbitrary number of classes in holistic scene understanding, *e.g.*, in the autonomous driving scenario regarding panoramic distortion, the performance of the model should be further improved for the safety of both drivers and pedestrians. Apart from the outdoor applications, the indoor navigation of robots using panoramas is supposed to focus more on open-vocabulary prediction correctness so that robots can better serve humans, avoiding unexpected potential dangers.

References

1. Kim, S., Kim, J., Kim, T., Heo, H., Kim, S., Lee, J., Kim, J.H.: Panoramic image-to-image translation. arXiv preprint arXiv:2304.04960 (2023)
2. Ray, B., Jung, J., Larabi, M.: A low-complexity video encoder for equirectangular projected 360 video content. In: ICASSP (2018)
3. Yang, K., Zhang, J., Reiß, S., Hu, X., Stiefelhagen, R.: Capturing omni-range context for omnidirectional segmentation. In: CVPR (2021)