Appendix of PanoFormer

Zhijie Shen[1,2], Chunyu Lin[1,2], Kang Liao[1,2,3], Lang Nie[1,2], Zishuo Zheng[1,2], and Yao Zhao[1,2]

[1] Institute of Information Science, Beijing Jiaotong University, China
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology
[3] Max Planck Institute for Informatics, Germany
Corresponding Author: Chunyu Lin
{zhjshen, cylin, kang_liao, nielang, zszheng, yzhao}@bjtu.edu.cn
Code: https://github.com/zhijieshen-bjtu/PanoFormer

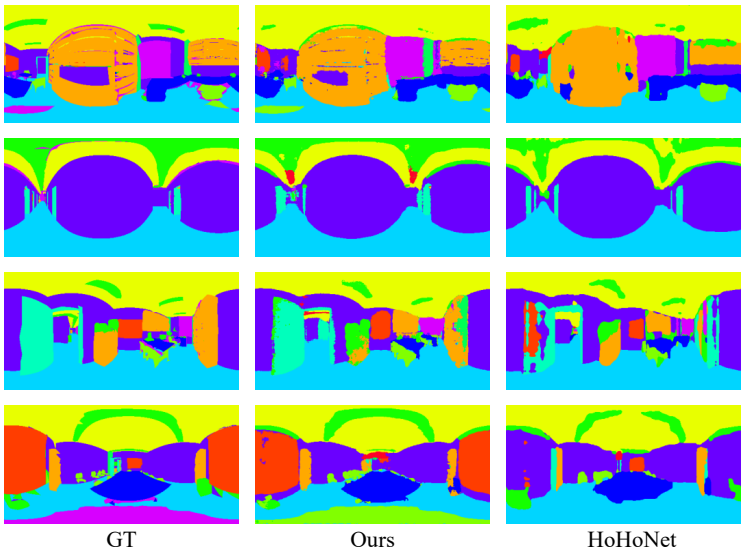

Fig. 1. Qualitative comparison between our approach and HoHoNet.

1 Panoramic Semantic Segmentation

Strictly following the experimental protocol described in [5], we validate the extensibility of our model on Stanford2D3D dataset with a resolution of 512×256. Figure 1 shows the qualitative comparison results between our model and HoHoNet [5]. We can observe that our results are more accurate and the details are finer. Tangent-patch dividing method for Panoformer significantly remove the negative effect of panoramic distortions, and further promotes token flow to perceive structure of objects. That help our network build a better panoramic perception capability. There are two ways to scale to high-resolution, 1) adding

Conv layers and 2) applying PanoFormer directly. However, PanoFormer is quadratic to the number of input pixels, so we recommend method 1). We run the PanoFormer using 1) and achieve mIoU at 56.5 with a 1024×2048 resolution.

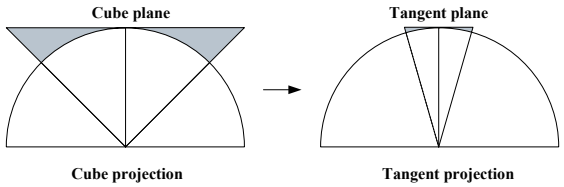

Fig. 2. Comparison with cube projection and tangent projection.

2 Analysis of Removing Distortion with Tangent Patch

The cube projection[2,3,4,6] is a commonly used representation of panoramic images, which represents a 360° scene by six cube faces (illustrated in Fig.3). Unlike the equirectangular map, its distortion gradually increases from the center of the projection surface to the periphery. We identify this change in Fig. 2 (the area in blue). It is easy to imagine that when the size of the cube plane is gradually reduced, the distortion around the projection surface will also decrease. A limit state is that there are only nine projection points on the projection surface, a central projection point, and eight surrounding projection points. In this case, the projection surface is approximately spherical, and the projection distortion is approximately zero. This limit state is our proposed tangent patch. We achieve tangent patch division in the ERP domain by the proposed relative position embedding method.

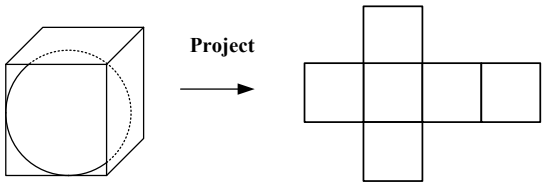

Fig. 3. Schematic of the cube projection.

3 Complexity

For a resolution of 1024×512: 80.91 GFlops, 14 fps; For a resolution of 512×256: 80.59 GFlops, 14 fps. In resolution 1024×512, we adjust the input/output stems due to the GPU memory limitation (stride set to 2 in input stem, adding interpolation layer in output stem; that's why the Gflops and fps are similar in different resolution).

Table 1. Results for necessity of PanoFormer.

Dataset	Method	MRE	MAE	RMSE	δ_1
Stanford-2D3D	normal patches	0.3162	0.4142	0.9585	0.6021
	'non-distortion' patches	Not converged			

4 Necessity of PanoFormer

We conducted two experiments: 1) Divide 16×16 patches directly on the ERP map, and predict the depth through the pre-trained ViT network. 2) Project ERPs into undistorted 16×16 patches; employ pretrained ViT network to predict depth; project the patches back to ERPs at the last layer. The results are shown in Table 1. Method 1) get a much worse result and method 2) does not converge. The reason is that the relative spatial positions among the projected 'non-distortion' tangent patches have been changed, making the network unaware of the actual positions in ERP. Besides, in [1], we found that ViT blocks are only used to implement patch embedding and the authors further design a Spatial Relationship Prediction to learn the relative spatial position of two tangent patches. This finding further confirms that the relative spatial position among tangent patches is hard to learn only by using ViT blocks. In PanoFormer, the panorama-customized spherical tokens, position embedding strategy, and token flows can all help to indicate the real positions.

References

1. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. Advances in Neural Information Processing Systems **34**, 5834–5847 (2021)
2. Jiang, H., Sheng, Z., Zhu, S., Dong, Z., Huang, R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. IEEE Robotics and Automation Letters **6**(2), 1519–1526 (2021)
3. Pearson, I.F.: Map Projections: Theory and Applications (1990)
4. Shen, Z., Lin, C., Nie, L., Liao, K., Zhao, Y.: Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
5. Sun, C., Sun, M., Chen, H.T.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2573–2582 (2021)
6. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 462–471 (2020)