Supplementary Material for : PanoMixSwap – Panorama Mixing via Structural Swapping for Indoor Scene Understanding

Yu-Cheng Hsieh
sphinx5912@gapp.nthu.edu.tw
Cheng Sun
chengsun@gapp.nthu.edu.tw
Suraj Dengale
surajdengale@gapp.nthu.edu.tw
Min Sun
sunmin@ee.nthu.edu.tw

Vision Science Lab
National Tsing Hua University
Hsinchu, Taiwan

1 PanoMixSwap vs. PanoStretch in Semantic Segmentation

Table 1 shows the comparison between our proposed method, PanoMixSwap, and the widely used panoramic augmentation technique PanoStretch proposed by Sun *et al*. [3]. We train HoHoNet [2] on Stanford2D3D [1] with 1024×2048 input resolution. Training and dataset settings are the same as those described in the main paper's Sec. 4.2. The results demonstrate that training with PanoMixSwap yields better results compared to training without PanoMixSwap, regardless of whether PanoStretch is used or not. While we find that using only PanoMixSwap outperforms using both PanoMixSwap and PanoStretch. We speculate the reason is that our PanoMixSwap already augments the layout structure, so the additional diversity by PanoStretch is limited since PanoStretch would make the images less realistic. Noted that we have made the layout estimation comparison between PanoMixSwap and PanoStretch in Sec. 4.3 of the main paper.

PanoMixSwap	PanoStretch	mIoU(%)	mACC(%)
-	-	52.00	65.00
✓	-	**56.02**	**67.43**
-	✓	53.63	65.06
✓	✓	55.91	67.03

Table 1: Quantitative comparison between two panoramic augmentations— PanoStretch and PanoMixSwap on semantic segmentation task.

2 More Visualization Results from PanoMixSwap

More Visualization Results from our augmentation, PanoMixSwap, are shown in Fig. 1. The proposed PanoMixSwap can produce indoor panoramic images with exceptional quality as well as diversity.

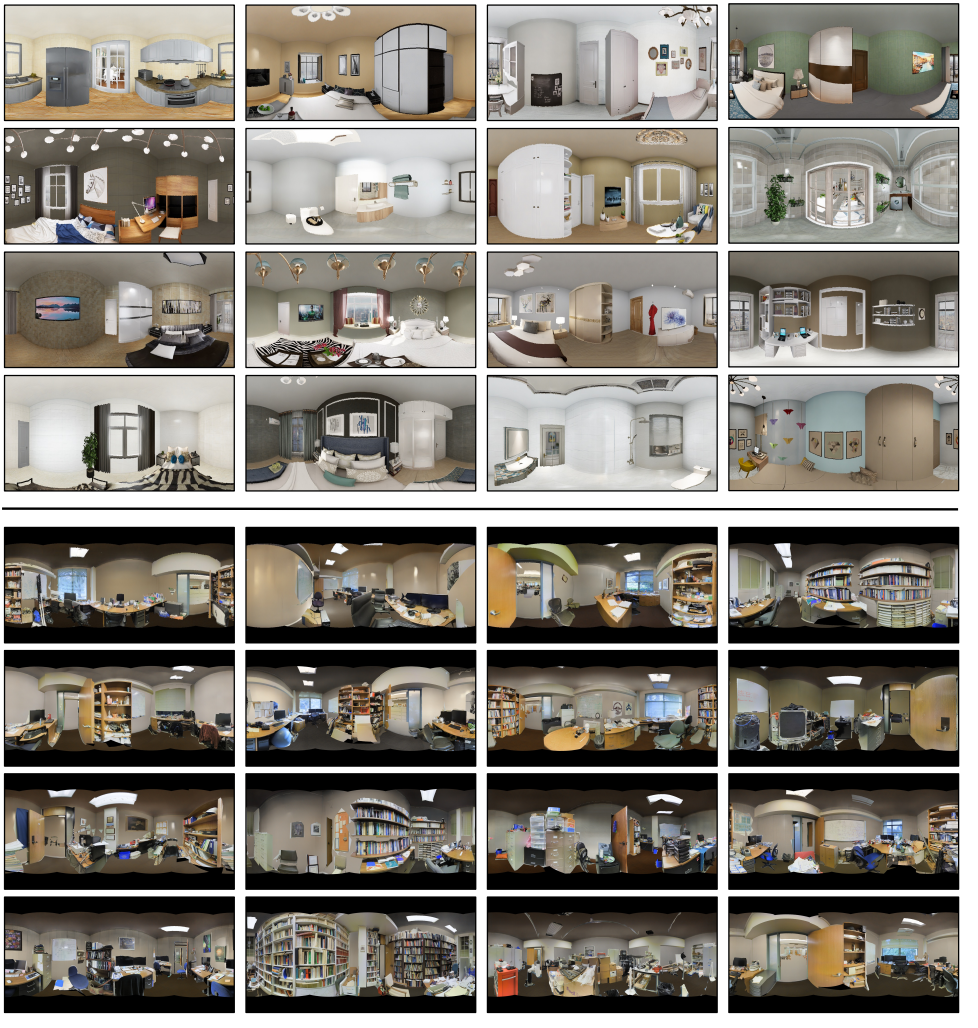

Figure 1: More visualization results of PanoMixSwap from Structured3D [5] (upper part) and Stanford2D3D [1] (lower part)

3 Qualitative Comparison for Semantic Segmentation

This section depicts the qualitative comparison between training with PanoMixSwap and with the original setting for the semantic segmentator HoHoNet [4] on Stanford2D3D [1] and Structured3D [5]. Qualitative results in Stanford2D3D [1] are shown in Fig. 2, while results in Structured3D [5] are shown in Fig. 3. Training with PanoMixSwap can generate more accurate semantic masks on both Stanford2D3D [1] and Structured3D [5].

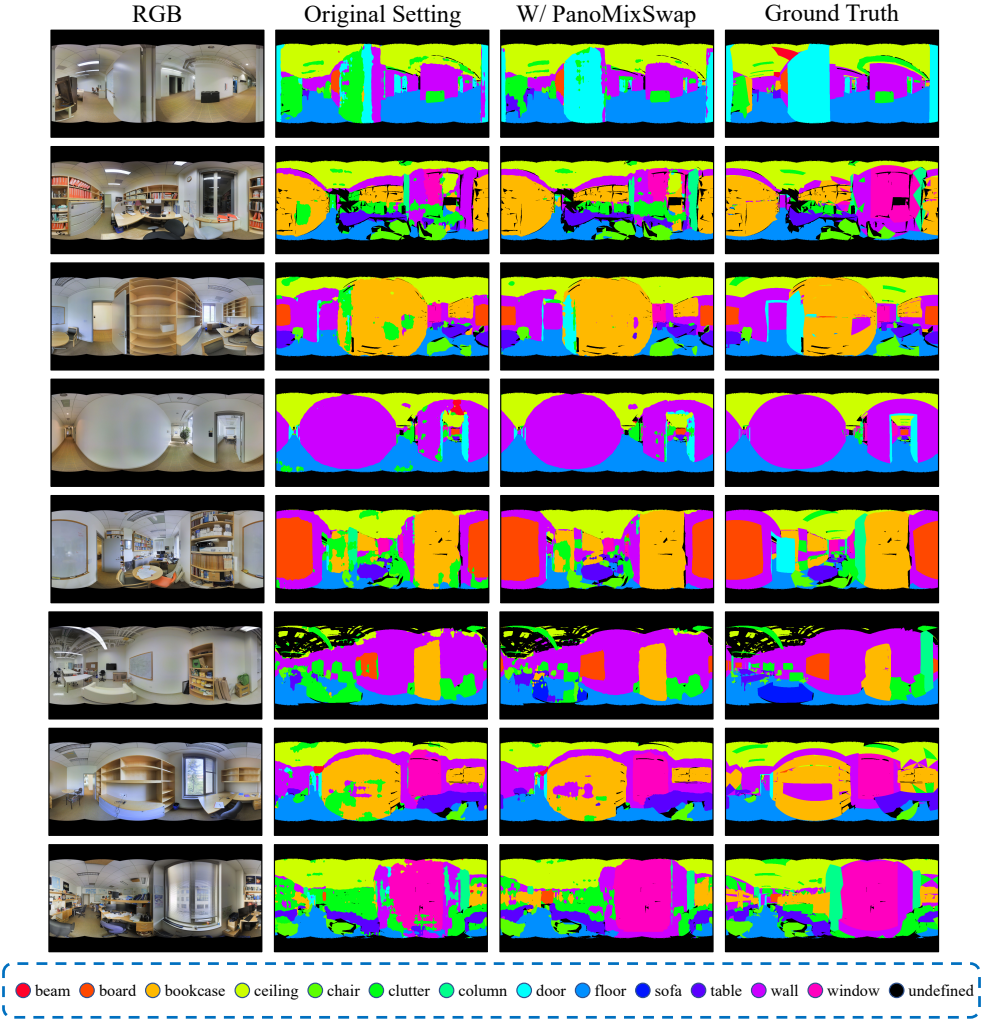

Figure 2: Qualitative comparisons for semantic segmentation on Stanford2D3D [1]. The results from the original setting column are obtained from the HoHoNet's [4] officially released pre-trained weights.

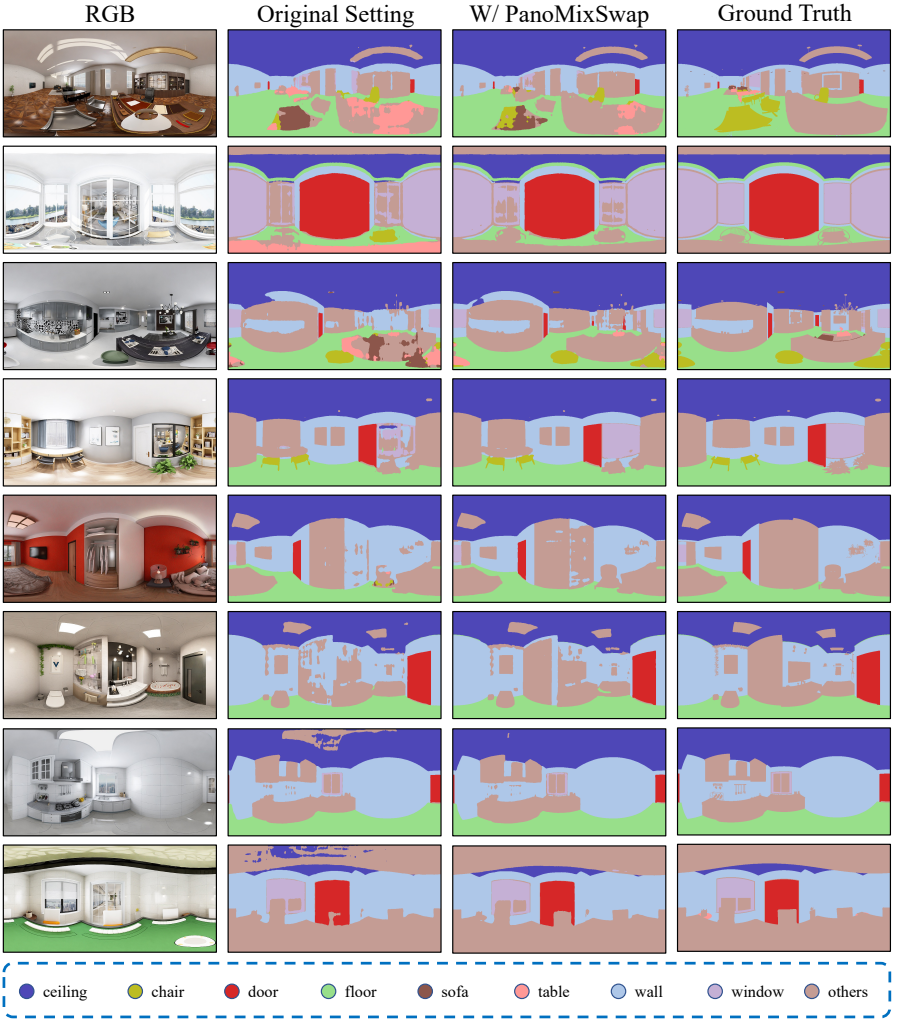

Figure 3: Qualitative comparisons for semantic segmentation on Structured3D [5]. For simplicity, we leverage HoHoNet [4] implementation on Structured3D [5], which reduces original 40 semantic classes to 9 classes.

4 Qualitative Comparison for Layout Estimation

Fig. 4 presents qualitative comparisons between training with PanoMixSwap and training with original settings for two layout estimators— HorizonNet [3] and LGT-Net [2] on the Stanford2D3D [1] dataset. Using PanoMixSwap for training can produce more accurate cuboid layouts compared to training with the original settings on both models, with HorizonNet [3] showing a particularly notable improvement. Noted that all original results of HorizonNet [3] and LGT-Net [2] are generated by their official pre-trained models.

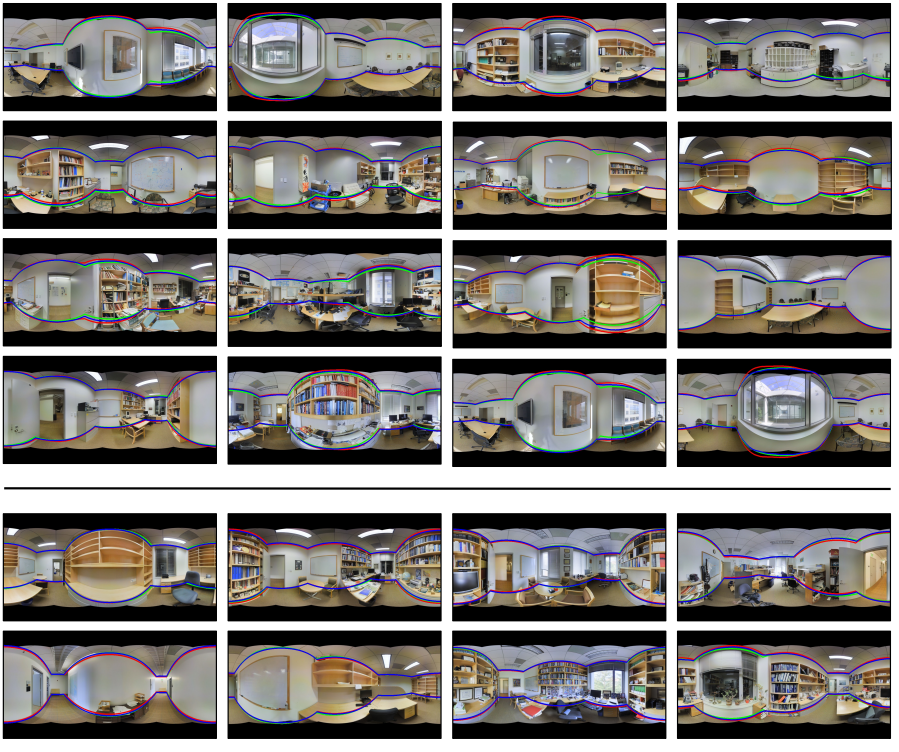

Figure 4: Qualitative comparisons of layout estimation on Stanford2D3D [1]. The results in the upper part are generated by HorizonNet [3] while those in the lower part are from LGT-Net [2]. The green, red and blue are the ground truth layout, original setting results and PanoMixSwap results.

References

[1] 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.

[2] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2022.

[3] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1047–1056, 2019.

[4] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021.

[5] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020.