MirageRoom: 3D Scene Segmentation with 2D Pre-trained Models by Mirage Projection

Supplementary Material

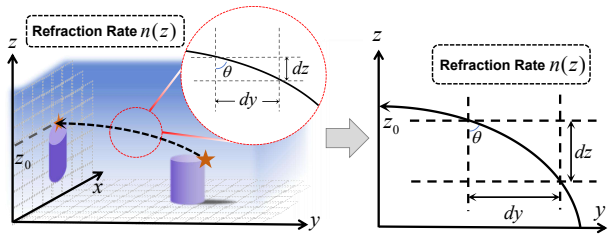


Figure 1. The illustration of projection ray in $y-z$ plane under the medium refraction rate distribution $n(z)$.

1. Detailed Derivation of Mirage Projection

As is illustrated in Figure 1, during the projection to $x-z$ plane, the distorted projection ray under heterogeneous distribution of medium is parallel to $y-z$ plane, thus the curve of ray can be formulated in $y-z$ coordinates. The typical medium refraction rate distribution $n(z)$ where mirage phenomenon occurs over sea [2, 3] can be formulated as:

$$n(z) = 1 + \rho_0 e^{-kz}, \tag{1}$$

which is only related to height z and has been introduced in Section 3.1 of main text. ρ_0 and k are constant coefficients which indicate the degree of change in refraction rate along z axis. According to optical principles, the angle of refraction (*i.e.*, θ in Figure 1) follows Snell's Law:

$$n(z)\sin\theta = C, \tag{2}$$

where C is a constant number, and θ is the angle between the orientation of ray and the vertical z axis, respectively. Since θ is an acute angle, (2) can be further formulated as:

$$\begin{aligned}\tan\theta &= \frac{\sin\theta}{\sqrt{1-\sin^2\theta}} \\ &= \frac{C/n(z)}{\sqrt{1-C^2/n^2(z)}} \\ &= \frac{C}{\sqrt{n^2(z)-C^2}} \\ &= \frac{C}{\sqrt{(1+\rho_0 e^{-kz})^2 - C^2}}.\end{aligned} \tag{3}$$

Meanwhile, the angle θ can be further represented as:

$$\tan\theta = \frac{dy}{dz}, \tag{4}$$

where dy and dz denote the differential of y and z, respectively. Combining (3) and (4), we can derive:

$$\begin{aligned}\frac{dz}{dy} &= \frac{1}{\tan\theta} \\ &= \sqrt{\frac{(1+\rho_0 e^{-kz})^2 - C^2}{C^2}} \\ &= \sqrt{\frac{(1+\rho_0 e^{-kz})^2}{C^2} - 1}.\end{aligned} \tag{5}$$

It is worth notice that (5) is *not* a standard differential equation of the curve since the signs of dz and dy in Figure 1 should be different. Moreover, we set $\hat{z} = z - z_{\text{max}}, \hat{z} \le 0$ where z_{max} is the boundary height of the ray to share the approximation of exponential function. Given the boundary condition where the projection ray far from ground (*i.e.*, $\hat{z} = 0$) is parallel to the ground (*i.e.*, $x-y$ plane), we can further formulate constant C as:

$$C = n(0)\sin(\pi/2) = 1 + \rho_0. \tag{6}$$

Therefore, we can formulate the differential equation of $\hat{z}$ as:

$$\frac{d\hat{z}}{dy} = -\sqrt{\frac{(1+\rho_0 e^{-k\hat{z}})^2}{(1+\rho_0)^2} - 1}. \tag{7}$$

The exact solution to (7) is a transcendental function, which is impractical for implementation. Therefore, we retain the second-order approximation concerning ρ_0 and k to simplify the expression since ρ_0, k are relatively small in real world:

$$\begin{aligned}(1+\rho_0 e^{-k\hat{z}})^2 &\approx (1+\rho_0-\rho_0 k\hat{z})^2 \\ &\approx (1+\rho_0)^2 - 2(1+\rho_0)\rho_0 k\hat{z} \\ &\approx (1+\rho_0)^2 - 2\rho_0 k\hat{z}.\end{aligned} \tag{8}$$

Based on the approximation, (7) can be further formulated as:

$$\begin{aligned}\frac{d\hat{z}}{dy} &= -\sqrt{\frac{(1+\rho_0)^2 - 2\rho_0 k\hat{z}}{(1+\rho_0)^2} - 1} \\ &= -\sqrt{\frac{-2\rho_0 k\hat{z}}{(1+\rho_0)^2}} \\ &= -\sqrt{\frac{2\rho_0 k}{(1+\rho_0)^2}}(-\hat{z})^{\frac{1}{2}}.\end{aligned} \tag{9}$$

It is clear that the analytical solution to (9) corresponds to a parabola. Considering the boundary condition where $\hat{z} = 0$

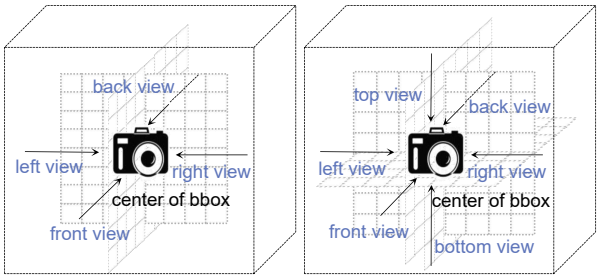


Figure 2. Different choices of projection planes. Both two choices set the center of the bounding box of input point cloud as the center of projection planes. (A) includes 4 view planes, *i.e.* front view, back view, left view and right view, while (B) includes 2 more views from top and bottom.

is attained at $y = 0$, we have:

$$\hat{z} = -\frac{\rho_0 k}{2(1+\rho_0)^2} y^2, \tag{10}$$

Therefore, assuming the boundary height where the ray meets view plane is z_0 (*i.e.*, z_0 in Figure 1), the equation can be further expressed as:

$$z = z_0 - \frac{\rho_0 k}{2(1+\rho_0)^2} y^2. \tag{11}$$

which shares the same representation with our final implementation in mirage projection.

2. More Implementation Details

In this section, we further provide more implementation details for data processing and data augmentations in training/testing process.

Data preprocessing. We follow the conventional practise in [7, 8] to preprocess two datasets. For S3DIS [1], we use the aligned version. Since the training set is relatively small, we follow [8, 9] to enlarge the size by repeating 30× to obtain 6,120 scenes. For ScanNet V2 [4], we also repeat 9× training set to get 10,809 training samples. We also collect the normal vector of each point.

Data augmentation in training/testing process. During training process, we adopt same data augmentation strategies with [8], including random dropout, random flip, random scale, random jittering. During training process, we use sphere crop over the entire scene and constrain the maximum number of input points to 100,000 after group mirage projection. We normalization the projection image with the same strategy in 2D image processing [5]. It is worth notice that our method does not require Mix3D [6] strategy which is applied in PTv2 [8]. During testing process, we use a test-time voting strategy which is the same with previous works.

Table 1. The comparison of different choices of projection planes in Figure 2 on S3DIS Area 5.

Choice of Projection Planes	n_v	mIoU (%)
(A)	4	72.0
(B)	6	71.3

Table 2. The comparison of fair data augmentation between PTv2 and our method on S3DIS Area 5.

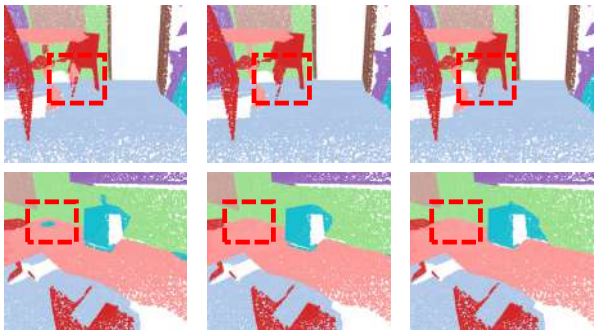
Method	Mix3D Aug	Max points per sample	mIoU (%)
PTv2 [8]	✓	30k	71.6
	✗	10k	71.0
MirageRoom	✗	10k	72.0

3. More Ablation Studies and Analyses

In this section, we further conduct ablation studies to verify the effectiveness of our experimental designs, including the choices of projection planes and the comparison under fair data augmentation.

Choices of projection planes. As is mentioned in Section 3.3 and 4.1 of main text, we use 4 projection planes to generate multi-view projection images for each κ, which is illustrated in Figure 2 (A). Specifically, we first set the center of the bounding box of input point cloud as the center of projection planes, then we choose 4 planes where the normal directions are front, back, left and right, respectively. We further add 2 more projection planes whose normal directions are top and bottom, which is illustrated in Figure 2 (B). In this way, the number of projection planes $n_v = 6$. The inference results of (A) and (B) on S3DIS area 5 are illustrated in Table 1. It is clear that the performance is better under (A), and two major reasons leads to the drop under (B): 1) Different distributions of objects between top/bottom views and other views, which makes the FPN network hard to generate unified feature representations. 2) Mirage projection is not suitable for top/bottom views due to the physical modeling.

Fair data augmentation. For a fairer comparison to evaluate the effectiveness of our method, we further train PTv2 [8] without Mix3D [6] data augmentation, which shares same augmentations with our method. The results are shown in Table 2. Apparently, the Mix3D strategy provides more points in one batch, and our method exhibits a



w/o Mirage Proj. w/ Mirage Proj. GroundTruth

Figure 3. Visualization results of our method compared with the same network architecture without group mirage projection but only straight-line projection. Best viewed in color.

more significant advantage over PTv2 under same data augmentation settings.

4. More Visualization Results

We visualize the detail results of our method compared with the same network architecture without group mirage projection but only straight-line projection, and the visualization results are shown in Figure 3. It is clear that our mirage projection can provide more accurate details even though it is occluded by straight-line projection. However, the experiment without group mirage projection makes mistakes.

More visualizations comparisons on S3DIS [1] and ScanNet V2 [4] between PTv2 [8] and our method are illustrated in Figure 4 and Figure 5, respectively. Compared with PTv2 [8], our method provides more accurate predictions especially over regions without clear geometric and structural features. For example, in the first 2 rows of Figure 4, PTv2 fails to predict murals on the wall, which shares similar geometric structures with wall. Thanks to the guidance from 2D projection, our method is able to obtain correct inferences for these parts.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, pages 1534–1543, 2016. 2, 3, 4

[2] Bradford R Bean and Gordon D Thayer. Models of the atmospheric radio refractive index. *Proceedings of the IRE*, 47(5): 740–755, 1959. 1

[3] Lamont V Blake. Ray height computation for a continuous nonlinear atmospheric refractive-index profile. *Radio Science*, 3(1):85–92, 1968. 1

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 3, 4

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[6] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. In *3DV*, pages 116–125. IEEE, 2021. 2

[7] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 30, 2017. 2

[8] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *NIPS*, 35:33330–33342, 2022. 2, 3, 4

[9] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pages 16259–16268, 2021. 2

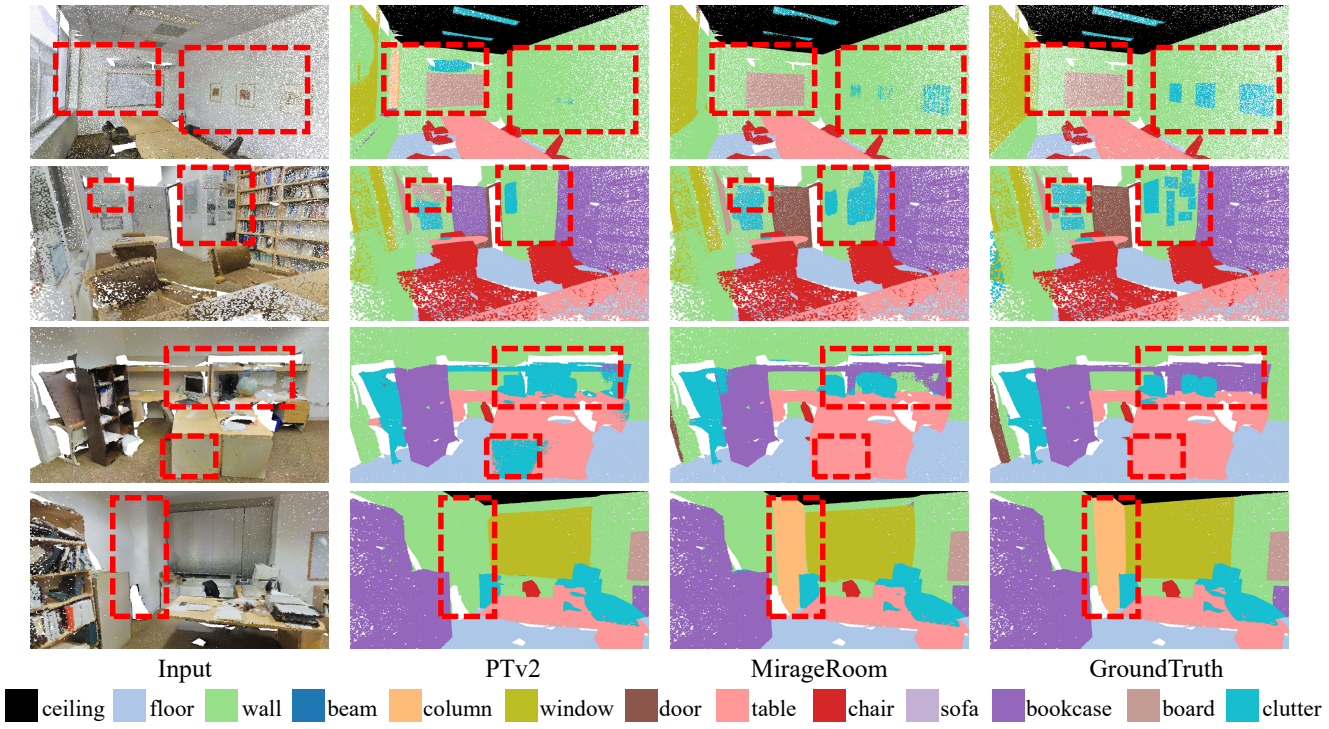


Figure 4. Visualization results of semantic segmentation results of PTv2 [8] and MirageRoom on S3DIS [1] dataset. Best viewed in color.

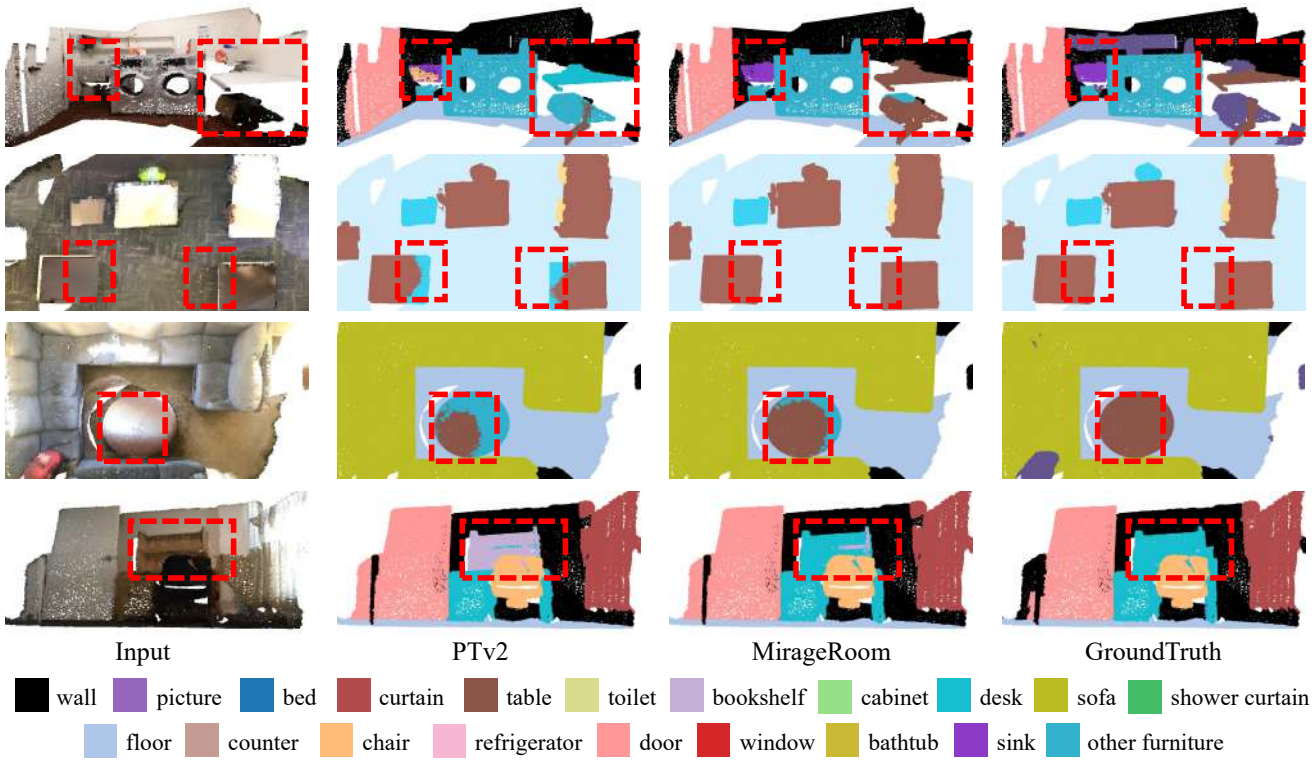


Figure 5. Visualization results of semantic segmentation results of PTv2 [8] and MirageRoom on ScanNet V2 [4] dataset. Best viewed in color.