Supplementary Material for Layout-Guided Novel View Synthesis from a Single Indoor Panorama

Jiale Xu[1] Jia Zheng[2] Yanyu Xu[3] Rui Tang[2] Shenghua Gao[1,4*]
[1]ShanghaiTech University [2]KooLab, Manycore
[3]Institute of High Performance Computing, A*STAR
[4]Shanghai Engineering Research Center of Intelligent Vision and Imaging
{xujl1, gaoshh}@shanghaitech.edu.cn {jiajia, ati}@qunhemail.com
xu_yanyu@ihpc.a-star.edu.sg

In this supplementary material, we first present the details of the spherical geometric transformation process. Then, we explain the relationship between pixel missing rate and the camera translation distance. Finally, we show more qualitative results and some failure cases.

1. Spherical Geometric Transformation

A description of the panorama. In this work, we assume that panoramas are generated with equirectangular projection, which means the horizontal and vertical pixel coordinates relate linearly to the longitude and latitude of the spherical coordinates, respectively. A panorama I covers $360°$ field-of-view (FoV) horizontally and $180°$ FoV vertically, and the center $(W/2, H/2)$ aligns to the 0-longitude and 0-latitude.

Different from perspective images, each pixel of the panorama corresponds to a direction in the spherical coordinate system. Its depth value denotes the 3D Euclidean distance from the camera center to the nearest point along this direction. Therefore, we can know the 3D location of each source-view pixel once the depth map $\boldsymbol{D}_s$ is estimated, then the feature map $\boldsymbol{F}_s$ can be seen as a point cloud $\boldsymbol{P}$ and projected into the novel views.

Mapping from source-view to target-view. The mapping from a source-view pixel $(u_s, v_s) \in \mathcal{P}_s$ to a target-view pixel $(u_t, v_t) \in \mathcal{P}_s$ is determined by a series of coordinate transformation f from $\mathcal{P}_s$ to $\mathcal{P}_t$:

$$f = f_{\mathcal{S}_t \mapsto \mathcal{P}_t} \circ f_{\mathcal{C}_t \mapsto \mathcal{S}_t} \circ f_{\mathcal{C}_s \mapsto \mathcal{C}_t} \circ f_{\mathcal{S}_s \mapsto \mathcal{C}_s} \circ f_{\mathcal{P}_s \mapsto \mathcal{S}_s}. \quad (1)$$

Given a source-view pixel $(u_s, v_s) \in \mathcal{P}_s$, we first transform it to spherical coordinates $(\phi_s, \theta_s) \in \mathcal{S}_s$ as follows:

$$f_{\mathcal{P}_s \mapsto \mathcal{S}_s} : \begin{cases} \phi_s = u_s \cdot 2\pi / W - \pi \\ \theta_s = -v_s \cdot \pi / H + \pi / 2 \end{cases} \quad (2)$$

The spherical coordinates (ϕ_s, θ_s) can only determine the viewing direction of the pixel, we still need its distance relative to the camera $\boldsymbol{p}_s$ to locate its 3D position. Fortunately, the distance d and can be obtained from the estimated depth map $\boldsymbol{D}_s$. Thus, the 3D position (x_s, y_s, z_s) of the pixel relative to camera $\boldsymbol{p}_s$ can be calculated by:

$$f_{\mathcal{S}_s \mapsto \mathcal{C}_s} : \begin{cases} x_s = d \cdot \sin(\phi_s) \cdot \cos(\theta_s) \\ y_s = d \cdot \cos(\phi_s) \cdot \cos(\theta_s) \\ z_s = d \cdot \sin(\theta_s) \end{cases} \quad (3)$$

Now we have a 3D point $(x_s, y_s, z_s) \in \mathcal{C}_s$, it is easy to transform it to the target view since there is only a translation $\boldsymbol{t} = [t_x, t_y, t_z]^T = \boldsymbol{p}_t - \boldsymbol{p}_s$ between the two views:

$$f_{\mathcal{C}_s \mapsto \mathcal{C}_t} : \begin{cases} x_t = x_s - t_x \\ y_t = y_s - t_y \\ z_t = z_s - t_z \end{cases} \quad (4)$$

Finally, we conduct the inverse mapping of Eq. (3) and Eq. (2) to project the 3D point $(x_t, y_t, z_t) \in \mathcal{C}_t$ to a target-view panorama pixel $(u_t, v_t) \in \mathcal{P}_t$:

$$f_{\mathcal{C}_t \mapsto \mathcal{S}_t} : \begin{cases} \phi_t = \arctan(x_t / y_t) \\ \theta_t = \arctan(z_t / \sqrt{x_t^2 + y_t^2}) \end{cases} \quad (5)$$

$$f_{\mathcal{S}_t \mapsto \mathcal{P}_t} : \begin{cases} u_t = (\pi + \phi_t) \cdot W / (2\pi) \\ v_t = (\pi / 2 - \theta_t) \cdot H / \pi \end{cases} \quad (6)$$

Details of layout transformation. The estimated 2D room layout $\boldsymbol{L}_s \in \mathbb{R}^{N \times 2}$ is an ordered list of the pixel coordinates of room corners. Two adjacent rows in $\boldsymbol{L}_s$ represent the upper and lower corners on the same wall-wall boundary, respectively.

The transformation from the source-view 2D layout $\boldsymbol{L}_s$ to the target-view 2D layout $\boldsymbol{L}_t$ is similar to the feature map transformation. For each 2D corner $(u_s, v_s) \in \mathcal{P}_s$, we

*Corresponding author.

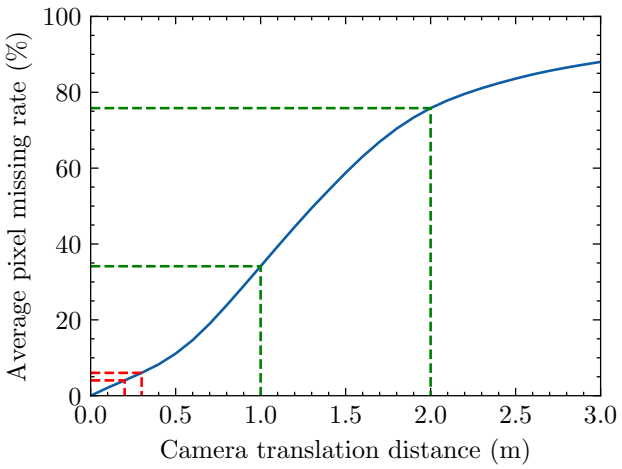


Figure 1. The relationship between the average pixel missing rate and the camera translation distance. The red dotted lines and the green dotted lines show the range of our generated easy set and hard set, respectively.

need to map it to $(u_t, v_t) \in \mathcal{P}_t$. The transformation from $(u_s, v_s) \in \mathcal{P}_s$ to $(\phi_s, \theta_s) \in \mathcal{S}_s$ is the same as Eq. (2). However, the corner depth d is not available from $\boldsymbol{D}_s$ since the corner may be occluded by foreground objects. Therefore, Eq. (3) cannot be applied directly.

Instead, we estimate the corner depth with the camera height h. Given the source camera position $\boldsymbol{p}_s = [p_x, p_y, p_z]$, the camera height relative to the ground is $h = p_z$. Let $(\phi_{s,u}, \theta_{s,u})$ and $(\phi_{s,l}, \theta_{s,l})$ denote the spherical coordinates of a pair of upper and lower corners, which belong to the same wall-wall boundary. We can estimate the depths of both corners as:

$$\begin{cases} d_u = h/|\tan(\theta_{s,l}) \cdot \cos(\theta_{s,u})| \\ d_l = h/|\sin(\theta_{s,l})| \end{cases} \tag{7}$$

With the estimated depths of all layout corners, we apply Eq. (3), Eq. (4), Eq. (5), and Eq. (6) to obtain their target-view pixel coordinates and form the target 2D layout $\boldsymbol{L}_t$.

2. More Details of Our Dataset

To clarify the difficulties of our dataset settings, we visualize the relationship between the pixel missing rate after the forward splatting operation and the camera translation distance. First, we randomly sample 100 panoramas from our dataset. Then we set the camera translation from $0\,\mathrm{m}$ to $3\,\mathrm{m}$ with a step size of $0.1\,\mathrm{m}$ in 5 random directions, and record the pixel missing rates after the forward splatting operation. As shown in Figure 1, the average pixel missing rate increases as the camera translation distance gets larger, which means that the difficulty of target-view inpainting rises. For the easy set, the camera translation ranges from $0.2\,\mathrm{m}$ to $0.3\,\mathrm{m}$, and the average pixel missing rate ranges from $4.04\,\%$ to $6.03\,\%$. For the hard set, the camera translation ranges from $1.0\,\mathrm{m}$ to $2.0\,\mathrm{m}$, and the average pixel missing rate ranges from $34.1\,\%$ to $75.8\,\%$, which is a very challenging setting that has rarely been considered before.

3. More Qualitative Results

More results on our synthetic dataset. Additional qualitative results on our synthetic easy set and hard set are shown in Figure 2 and Figure 3, respectively.

More results on real-scene datasets. Additional qualitative results on real-scene datasets, *i.e.*, 2D-3D-S dataset [1] and PanoContext dataset [2], are shown in Figure 4 and Figure 5, respectively.

Qualitative comparisons of the ablation study. Additional qualitative comparisons for the ablation study of the layout guidance are shown in Figure 6. We make the following observations: (i) both models generate plausible results within the seen regions. (ii) the model with layout guidance can preserve the room layout better. In some cases, the network can even generate unseen room structures from scratch with the help of the layout prior. (iii) the model without layout guidance produces blurry textures and indistinguishable layout boundaries in the unseen regions. The results demonstrate the effectiveness of the layout guidance.

4. Failure Cases

We show some typical failure cases in Figure 7. In Figure 7a and Figure 7b, the textures within the mirror and glass regions are incorrect. Since the depth estimation module does not explicitly model the reflection and refraction, the depth of the objects reflected by the mirror or behind the glass cannot be represented by the estimated depth map. In Figure 7c, some furniture in the scene has complex and thin structures, which are hard to be recovered from the coarse estimated depth map and result in conspicuous pixel distortion in the synthesized image. In Figure 7d, when the viewpoint change is too violent, *e.g.*, from one side to another side of an object, the network could not synthesize the unseen surface without any contextual information of the object.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 2

[2] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV*, pages 668–686, 2014. 2

Figure 2. Additional qualitative results on our synthetic easy set.

Figure 3. Additional qualitative results on our synthetic hard set.

Source View | Camera Translation = 0.5 m | Camera Translation = 1.0 m | Camera Translation = 1.5 m

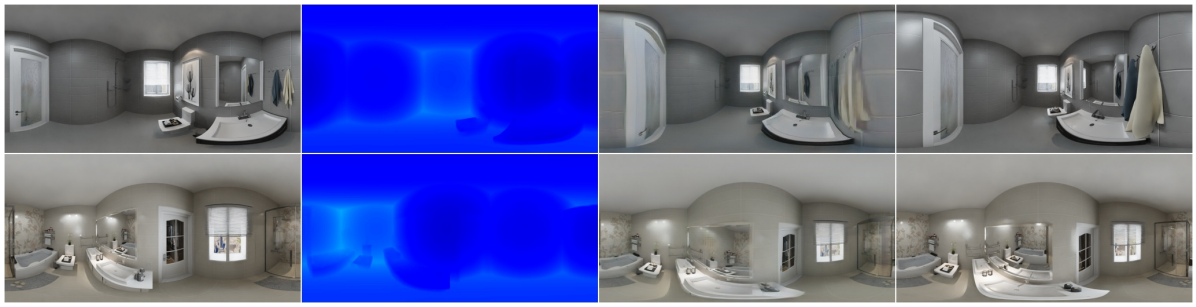
Figure 4. Additional qualitative results on 2D-3D-S dataset.

Source View | Camera Translation = 0.5 m | Camera Translation = 1.0 m | Camera Translation = 1.5 m

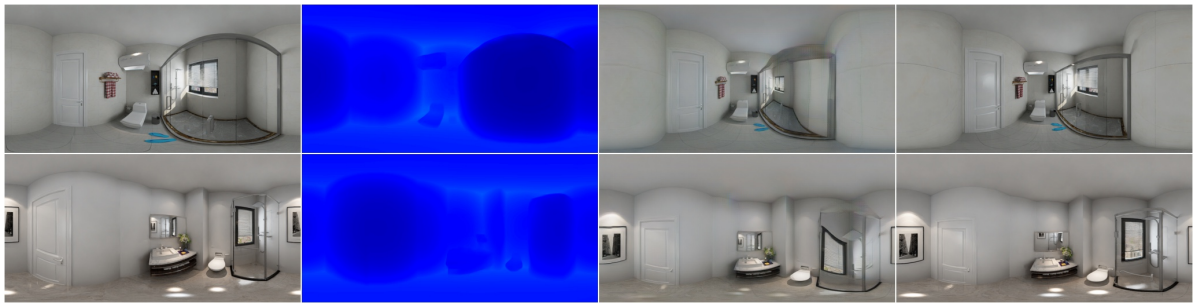
Figure 5. Additional qualitative results on PanoContext dataset.

Source View | Ours (without layout) | Ours (with layout) | Target View (Ground Truth)

Figure 6. Additional qualitative comparisons for the ablation study of the room layout guidance.

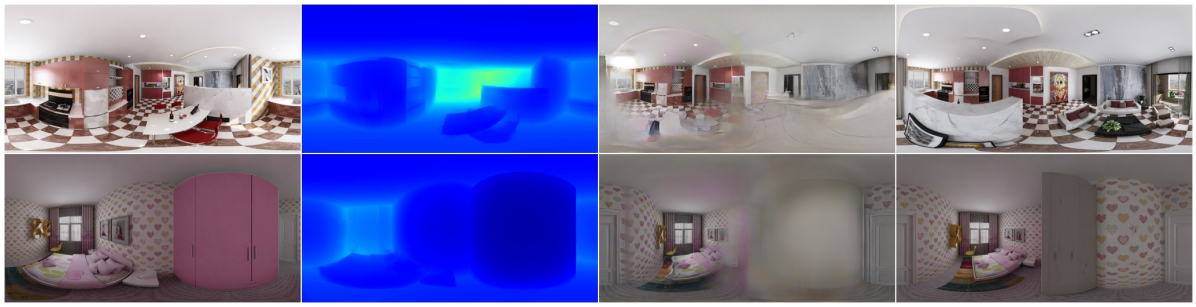


(a) Mirrors.



(b) Glass.

(c) Fine structures.



(d) Lack of context.

Figure 7. Failure cases on our synthetic dataset. From left to right: source-view image, estimated depth map, synthesized target-view image, and ground-truth target-view image.