Scene4U: Hierarchical Layered 3D Scene Reconstruction from Single Panoramic Image for Your Immerse Exploration

Supplementary Material

In this supplementary material, we provide additional details that could not be included in the manuscript due to space limitations. Specifically, we address the following aspects:

- Overview of WorldVista3D dataset.
- More Implementations Details.
  - Configuration.
  - Model Setting.
- More Experiment Results.
  - Results of Climate Controller.
  - More Scene4U Results.

1. Overview of WorldVista3D dataset

An overview of the WorldVista3D dataset used in this paper is provided in Fig. 1, highlights the global distribution of the panoramic images included. The panoramic images span five continents and capture a diverse array of landmarks with cultural, historical, and architectural significance. Examples include Stonehenge and the Louvre in Europe, 1the Great Wall in Asia, and the Sydney Opera House in Australia. The WorldVista3D dataset provides extensive opportunities for 3D scene reconstruction from a single panoramic image.

Figure 1. **Overview of panorama image location in WorldVista3D**.

2. More Implementations Details

2.1. Configuration

We provide a detailed explanation of the experimental environment and the specific optimization parameter settings:

- **Environment Setting**
  - Platform: Linux
  - CUDA version: 11.8
  - PyTorch version: 2.0.1
  - GPU: NVIDIA A100 Tensor 80GB
- **3D Gaussian Splatting Optimization parameters**
  - position_lr_init = 0.00016
  - feature_lr = 0.0025
  - opacity_lr = 0.06
  - scaling_lr = 0.005
  - rotation_lr = 0.001
  - opacity_reset_interval = 3000

2.2. Model Setting

Multi-Layer Segmentation. We utilize the pre-trained model within the open-vocabulary segmentation framework, Semantic Segment Anything[2], to segment objects in a 2048×1024 resolution panorama P_0. The segmentation results are subsequently processed by an LLM[1], which classifies all objects in four layers: the dynamic object layer M_{dynamic} (e.g., pedestrians, vehicles), the foreground layer $M_{\text{foreground}}$ (e.g., traffic lights, bollards, utility poles), the background layer $M_{\text{background}}$ (e.g., buildings, forests), and the sky layer M_{sky} (e.g., sun, clouds), based on the semantic segmentation results.

Multi-layer Image Restoration. We select the FLUX-inpainting model, which is trained on 1024×1024 high-resolution images, to provide high-quality and detailed inpainting results. Our goal is to use the inpainting model to restore specific masked regions of the panorama after it has been segmented into layers. First, we apply the dynamic object mask to remove dynamic objects, producing a serene version of the panorama, $P_{foreground}$, free of dynamic elements. Next, $P_{foreground}$ serves as the base for restoring the panorama, resulting in the background layer restoration map $P_{background}$. Finally, the $(1 - M_{sky})$ mask is used to restore the sky layer, producing the panoramic sky layer map, P_{sky}. The restoration parameters are as follows: the input panorama resolution is 2048×1024, and the inference steps are set to 30. For restoring the foreground and background layers, the prompt is configured as *empty scene, nothing, without any people and vehicles*, and for the sky layer restoration, the prompt is set to *hdri-360, sunny, clear sky with cloud, skybox, only sky, outdoors, HDR, empty scene, high altitude, no ground*.

Multi-layer Depth Completion. We employ the depth restoration model from Infusion[3], as shown in Fig. 2, which uses the depth map, mask, and color image to complete the depth information based on pixel correspondence between the color image and the mask. Specifically, we per-

form depth estimation on the $P_{\text{foreground}}$ to obtain $D_{\text{foreground}}$. Subsequently, $(P_{\text{background}}, D_{\text{foreground}}, M_{\text{foreground}})$ are together input into the depth completion model to generate the background depth completion result $D_{\text{background}}$. Finally, $(P_{\text{sky}}, D_{\text{background}}, 1 - M_{\text{sky}})$ are processed to produce the depth completion result for the sky layer D_{sky}.

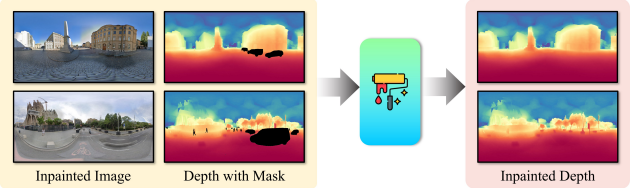

Figure 2. **Illustration of Depth Completion**. With multi-layer panoramas, depth maps, and masks, the model restores occluded structures and generates unoccluded depth for 3D reconstruction.

3. More Experiment Results

In this section, we exhibit the role of the Climate Controller module in our Scene4U and more qualitative results in various real-world scenes.

3.1. Results of Climate Controller

The visualizations of the Climate Controller are shown in the Fig.3. The Climate Controller is able to effectively generate scene environments for different seasons and weather conditions based on real panoramic images, offering users a variety of real-world scene options.

3.2. More Scene4U Results

Fig.4 shows the immersive 3D scenes of world-renowned landmarks provided for users. We present the layered images and depth, highlighted with frames in different colors, along with the final multi-view rendering results. The proposed method leverages a layered strategy to generate 3D scenes. It effectively mitigates issues related to scene noise and occlusion, thereby providing users with a more immersive experience.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[2] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023. 1

[3] Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024. 1

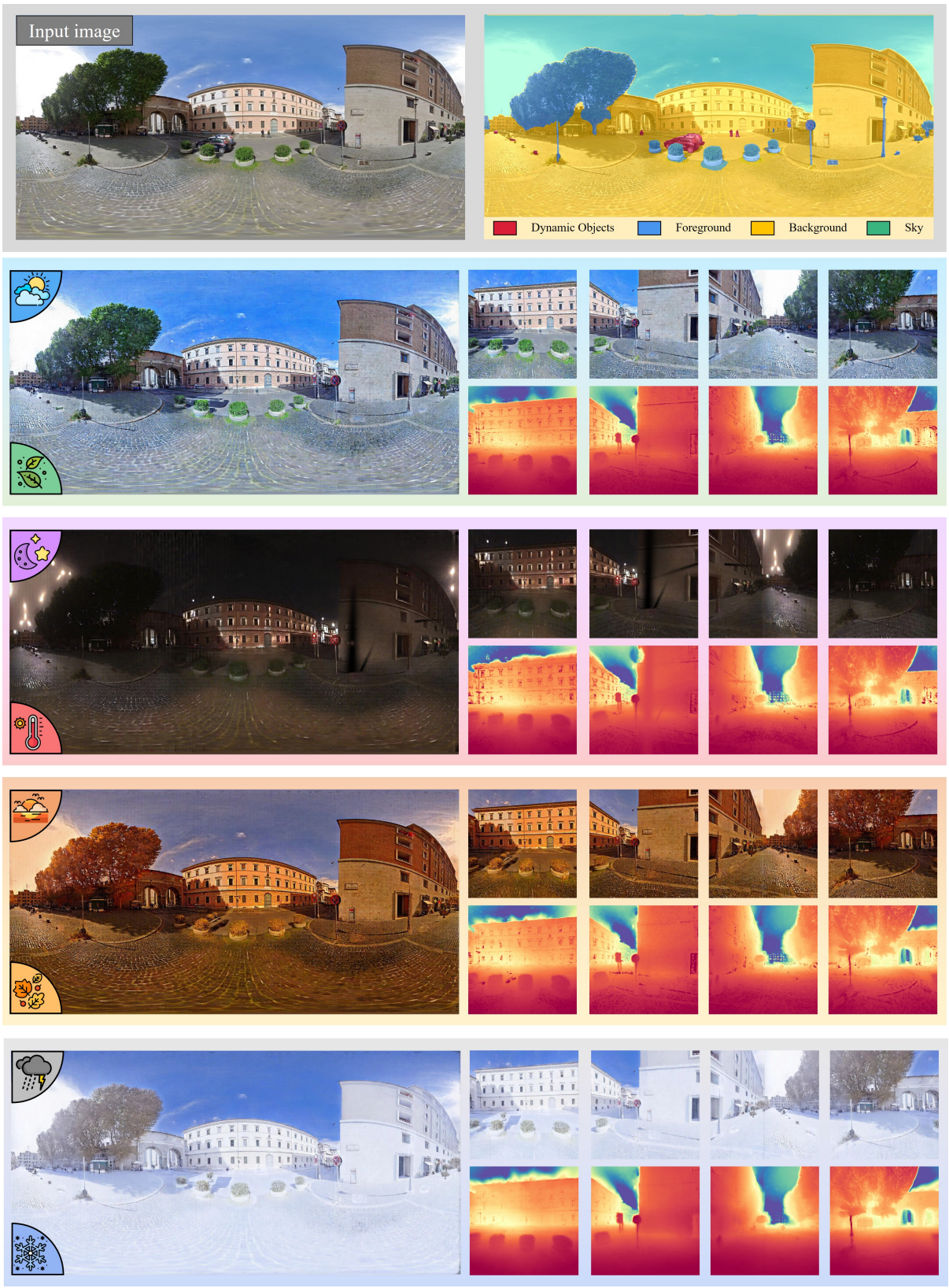

Figure 3. **Qualitative results of Climate Controller.** We present the scene reconstruction results under four temporal states: (spring, sunny), (summer, evening), (autumn, sunset), and (winter, rainy). With the assistance of Climate Controller, users can freely select and navigate through scenes under various temporal states.

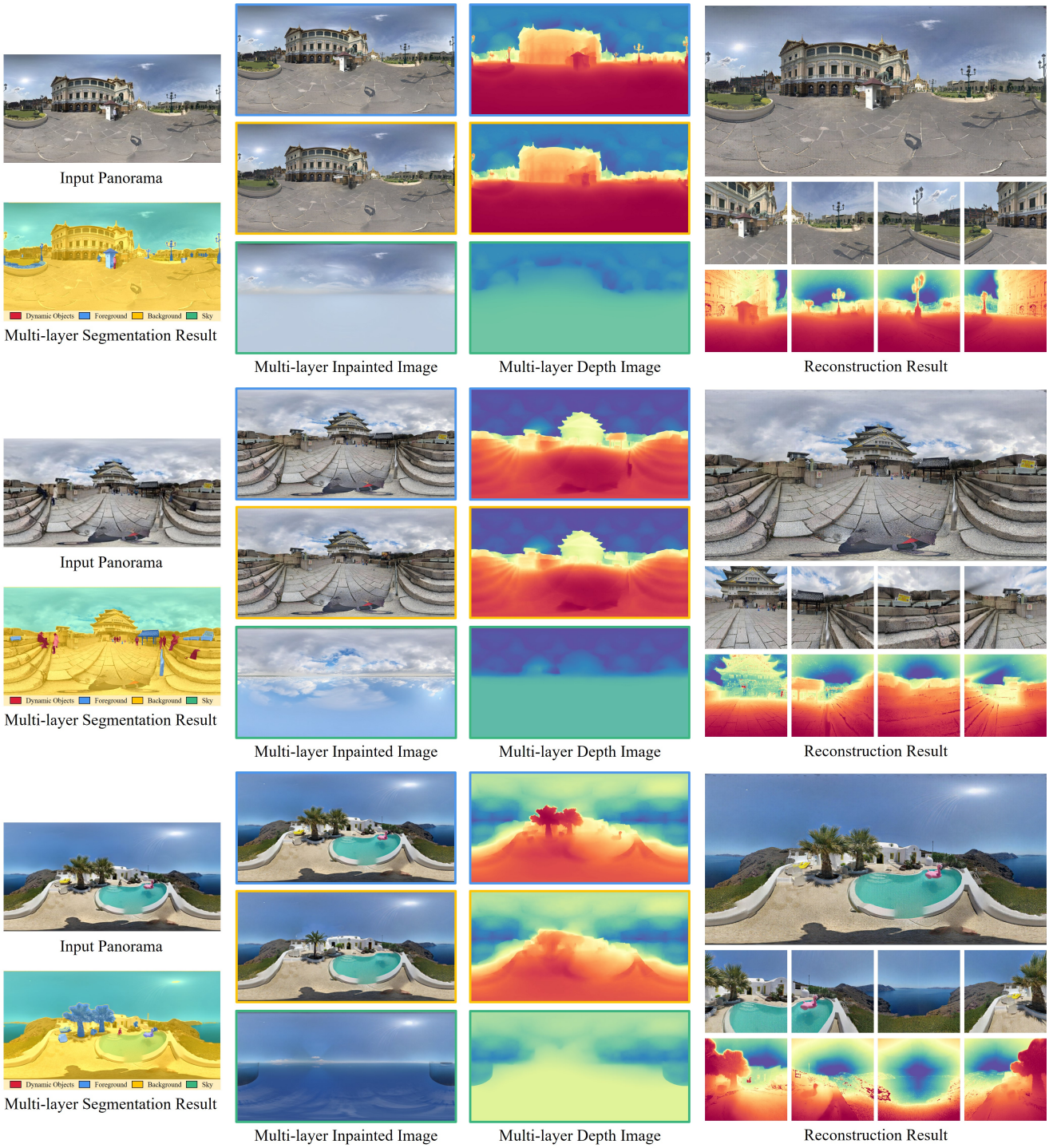

Figure 4. **Qualitative results of Scene4U.** We present the intermediate results, where the border colors of the multi-layer restoration and depth results correspond to the colors in the multi-layer segmentation map: blue borders indicate the foreground layer, yellow borders indicate the background layer, and green borders indicate the sky layer.

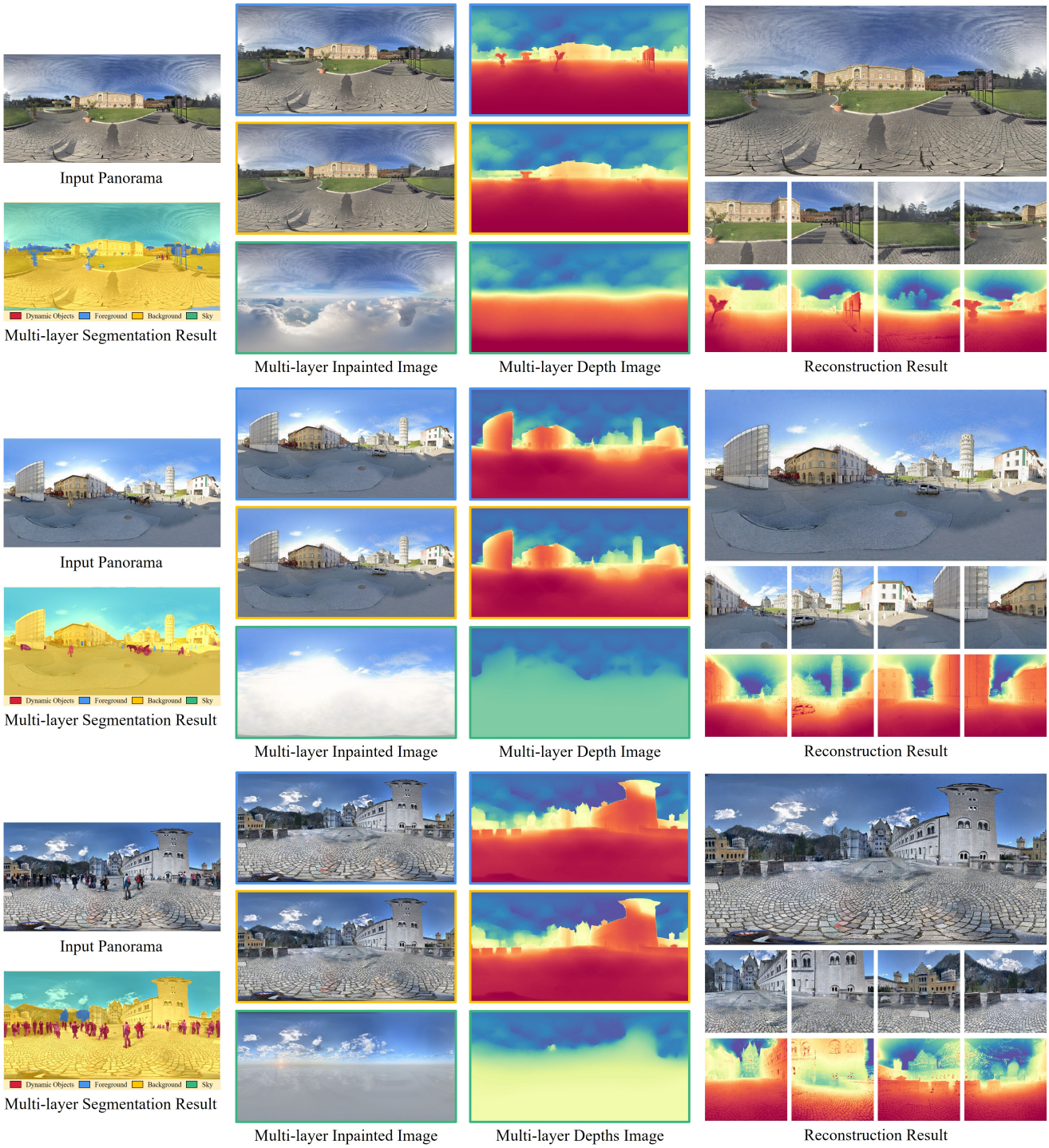

Figure 4. **Qualitative results of Scene4U.** We present the intermediate results, where the border colors of the multi-layer restoration and depth results correspond to the colors in the multi-layer segmentation map: blue borders indicate the foreground layer, yellow borders indicate the background layer, and green borders indicate the sky layer.