# Supplementary Materials for MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond

Yixuan Li[1*] , Lihan Jiang[2*] , Linning Xu[1], Yuanbo Xiangli[1]
Zhenzhi Wang[1], Dahua Lin[1,2], Bo Dai[2✉]
[1] The Chinese University of Hong Kong [2] Shanghai AI Laboratory
ly122@ie.cuhk.edu.hk jianglihan@pjlab.org.cn
{xl020,xy019,wz122,dhlin}@ie.cuhk.edu.hk daibo@pjlab.org.cn

## 1. Method Details

**NeRF** [4] is the pioneering method that represents a scene implicitly using a fully-connected MLP network. It can synthesize high-quality novel views through volume rendering. We utilize 12 layers with the width 512 for both the coarse and fine networks. The 3D location of the maximum frequency for positional encoding is set to 16 instead of 10. Additionally, we incorporate three skip connections that concatenate the input to the activations of the 4th, 8th, and 10th layers.

**DVGO** [6] represents the scene by means of optimizing dense grids, comprising of density grids for scene geometry and feature grids with a shallow MLP network for view-dependent appearance in detail. We removes the coarse geometry search step because the view-count-based learning rate prior is ineffective for urban scenes. This is because the edge views are often underrepresented, making it challenging to optimize the algorithm. Furthermore, we set the number of voxels in the fine stage to $500^3$ to ensure a fair comparison with other methods.

**Instant-NGP** [5] introduces a multi-resolution hash encoding structure, which guarantees both efficiency and accuracy in the modeling process. We increased the number of parameters for the hash encoding. Specifically, we utilized 16 levels for the hash encoding, with each individual hash table containing $2^{22}$ entries. The resolution ranged from 16 to 65536.

**TensoRF** [2] models the scenes as a 4D tensor, consisting of 3D voxel grids and a multi-channel feature. These tensors are then decomposed into compact vector and matrix factors for more efficient processing. In our paper, we adapt the grid resolution to $500^3$ to better model the large-scale scene and upsample the grids resolution at the 2000th, 3000th, 4000th, 5500th, 7000th, 10000th, 12000th, 14000th iterations. Additionally, we discover that the softplus activation function is unstable for large-scale urban scenes and thus replace it with ReLU.

**MipNeRF-360** [1] employs a non-linear parameterization technique, coarse-to-fine online distillation, and a distortion loss function to address unbounded artifacts. We normalize the camera poses of the trained images into a unit sphere for the contraction operation. We utilize 4 layers-MLP with the width 256 for the propose network and 8 layers-MLP with the width 1024 for the nerf network.

For all the experiments in our paper, we resize the 1080P images into $540 \times 960$ for training and testing. Each single model except MipNeRF-360 is trained on a single Nvidia A100 GPU device for around 0.5-30 hours. We use 4 Nvidia A100 GPU to train MipNeRF-360 for around 10 hours.

## 2. More Results on Aerial Data

To investigate the characteristics of the grid-based methods and MLP-based nerf methods for large area modeling, we conducted experiments using aerial data of an entire small city. We increase the capacity of the MipNeRF-360 by extending its nerf network to 12 layers, and enlarge the grid resolution of TensoRF to $1500^3$. As shown in Table 2, the MLP-based nerf methods, NeRF and MipNeRF-360, exhibit a significant decline in performance when modeling larger areas, although the model has been enlarged. Unlike MLP-based methods that utilize continuous networks, the grid-based approach of TensoRF and Instant-NGP utilizes discrete grids. By increasing the grid resolution, the model exhibits almost no drop in performance when modeling larger areas. As illustrated in Figure 1, the learning performance for details significantly diminished for both NeRF and MipNeRF-360 after expanding the area size, despite the accompanying increase in model size. For TensoRF and Instant-NGP, expanding the area size with increasing the grid resolution in proportion does not significantly affect

| Block | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Height | 150 m | 150 m | 300 m | 500 m | 450 m | 450 m | 350 m | 350 m | 250 m | 200m |

Table 1: Heights of different aerial block splits for the *Small City* data collection.

| Data Type | **NeRF** [4] | | | **TensoRF** [2] | | | **Instant-NGP** [5] | | | **MipNeRF-360** [1] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Multi-model | 22.97 | 0.589 | 0.548 | 25.13 | 0.762 | 0.396 | 25.69 | **0.800** | **0.326** | **25.73** | 0.779 | 0.360 |
| Single-model | 21.69 | 0.510 | 0.638 | 25.09 | **0.752** | **0.370** | **25.46** | 0.751 | 0.402 | 24.41 | 0.689 | 0.469 |

Table 2: Performance comparison of different kinds of methods on the aerial data of our $MatrixCity$ benchmark with the multi-model and the single model settings. Note that the multi-model represents that we divide the small city into five blocks and train a separate model for each block. The single-model represents that we train a single model for the whole small city.

the learning of details. However, as shown int the second row of Figure 1, training high-rise buildings and low-rise buildings together with a shared bbox can lead to the air above the low-rise buildings being relatively dirty, which is a point that needs to be optimized in future work.

## 3. More results on Stree Data

We notice that BlockNeRF [7] uses a small block size ($\sim 0.031km^2$), so we experiment on a similar setting in Table 3. We also ablate block size and model capacity on street views in the table below. Note that the grid resolution of TensoRF is $300/500/800$, and the network width of MipNeRF-360 is $256/512/1024$ for small/medium/large model size. We use the same number of rendering samples as the original paper for all methods.

| Blocks | Size | Model | **TensoRF** | | | **MipNeRF-360** | | |
|---|---|---|---|---|---|---|---|---|
| | | | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| 1 | $0.540km^2$ | Large | 21.23 | 0.663 | 0.556 | 22.00 | 0.717 | 0.488 |
| 4 | $0.135km^2$ | Large | 22.14 | 0.710 | 0.472 | 25.83 | 0.802 | 0.381 |
| 8 | $0.068km^2$ | Large | 23.47 | 0.751 | 0.432 | 26.44 | 0.835 | 0.320 |
| 16 | $0.034km^2$ | Large | 24.24 | 0.798 | 0.376 | **26.67** | **0.858** | **0.274** |
| 16 | $0.034km^2$ | Medium | 23.60 | 0.767 | 0.417 | 26.03 | 0.822 | 0.345 |
| 16 | $0.034km^2$ | Small | 22.96 | 0.727 | 0.469 | 24.64 | 0.763 | 0.432 |

Table 3: Ablation on block size and model capacity on the street data of Small City.

## 4. Extension Study

To delve deeper into the challenges of novel view synthesis in the real-world urban scenes, we design two sets of experiments that involved changing lighting conditions and dynamic scenes .

### 4.1. Illumination

In the real-world scenes, the intensity and direction of the light change throughout the day. We decouple these two dimensions to explore the challenges involved. For the light direction, we collect data with different light angle from 0 degree to 90 degree with a interval of 5 degree. The first row of Figure 2 shows the interpolation result of the direction of light from the first image (0 degree) to the last image (90 degree). We can see that the shadow only appears lighter, but it does not capture the essential relationship between the interaction of light and the building structure. For light intensity, we collect data with three different light intensity. In the second row of Figure 2, we can observe a reasonable and continuous change from the first image to the last image. However, there still exists a visual discrepancy between the interpolated images and the ground-truth images. This implies that the decoupling of light intensity and scene's color is not well executed.

### 4.2. Dynamic

We gather a collection of street view images featuring moving people and traffic, aiming to model the stationary buildings and filter out the dynamic traffic using NeRF-W [3], a method that can distinguish between dynamic objects and static buildings. However, as shown in Figure 3, we encounter a setback during the decomposing process; some stationary objects such as parked cars and streetlights were accidentally filtered out, which is unacceptable.

## References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5460–5469. IEEE, 2022. 1, 2

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV (32)*, volume 13692 of *Lecture Notes in Computer Science*, pages 333–350. Springer, 2022. 1, 2

[3] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219. IEEE, 2021. 2

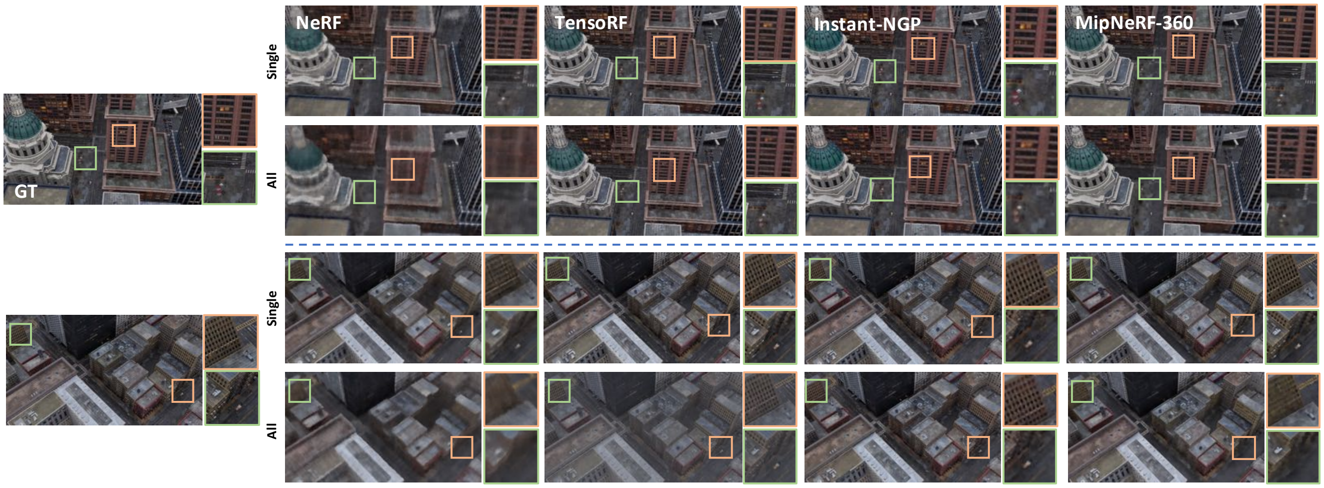

Figure 1: Visualization of novel view synthesis results of different kinds of methods on the aerial data of our $MatrixCity$ benchmark with the multi-model and the single model settings. Note that the multi-model represents that we divide the small city into five blocks and train a separate model for each block. The single-model represents that we train a single model for the whole small city.

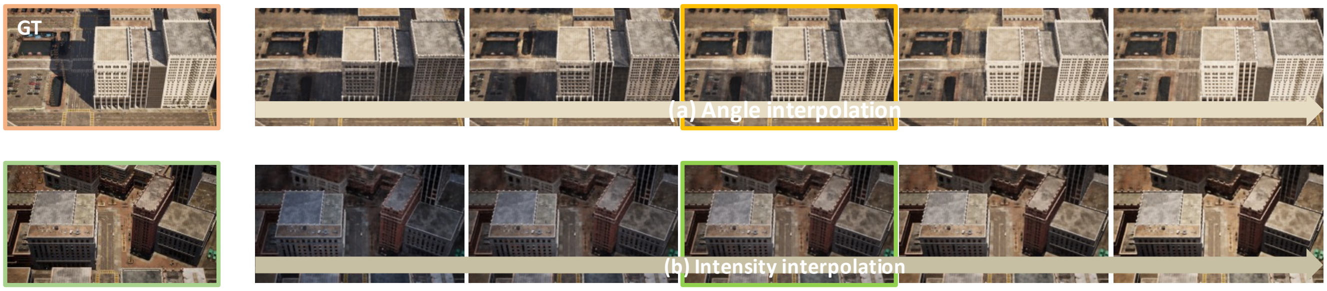

Figure 2: Visualization of the interpolation result of light angle (1st row) and light intensity (2nd row). In the first row, the angle of the first image is 0 degree and the angle of the last image is 90 degree. The middle image in the orange box is the interpolated 45 degree's result. In the second row, the intensity of the first image is 1000 and the intensity of the last image is 3000. The middle image in the green box is the interpolated 2000 intensity's result.

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer, 2020. 1, 2

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2

[6] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5449–5459. IEEE, 2022. 1

[7] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, pages 8238–8248. IEEE, 2022. 2

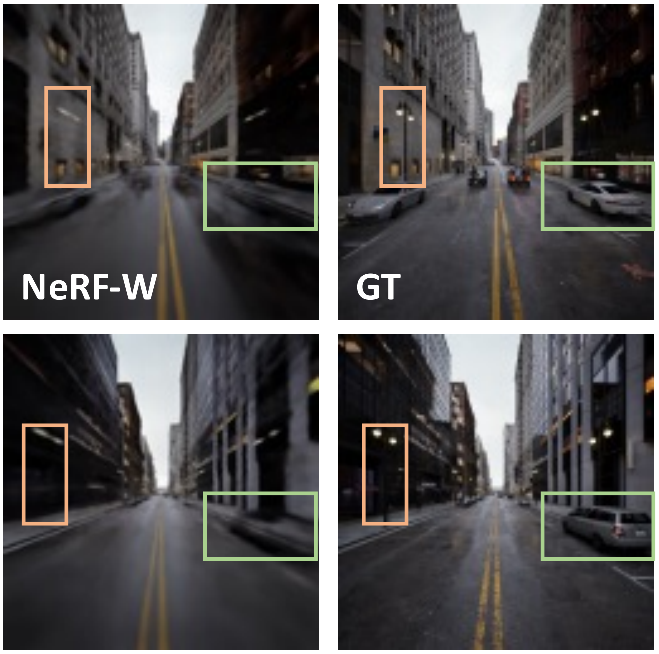

Figure 3: Visualization of the novel view synthesis results of the stationary scene with moving cars and people.