MagicCity: Geometry-Aware 3D City Generation from Satellite Imagery with Multi-View Consistency

Supplementary Material

6. Limitations and Discussion

6.1. Limitations

Although MagicCity produces photo-realistic 3D cities with geometric consistency and outperforms existing methods, there are still several limitations.

Geometric Diversity. Our method lifts voxels to specific heights based on the height map, meaning each pixel (x, y) corresponds to a single height value z. This limits surface detail and prevents complex structures like tunnels.

Edge Effects. Since our approach synthesizes multi-view images before 3D reconstruction, buildings near scene boundaries may be incomplete due to insufficient multi-view coverage, causing geometric artifacts.

Dependence on Semantic and Depth Estimation. The accuracy of our 3D city reconstruction relies on precise semantic segmentation and depth estimation. Errors in these steps can lead to structural inconsistencies.

6.2. Future direction

Our current approach uses satellite images as input, allowing us to directly convert them into virtual 3D cities, significantly reducing the cost of manual scene layout design. However, this also limits user customization. In the future, we plan to integrate large language models (LLMs) to enrich the input. This will enable users to specify city styles, building density, and the distribution of street-level objects through textual descriptions. Moreover, our 3D city generation currently operates at the block level due to computational constraints. In the future, we will optimize efficiency to scale up and generate larger urban areas. Additionally, our current method generates 3D city regions in a single pass. Moving forward, we aim to develop an interactive editing framework that allows users to make real-time modifications to individual assets after generation. Beyond these improvements, we will also address existing limitations such as geometric diversity and edge effects to further enhance the quality and flexibility of generated urban scenes.

7. Application scenarios

After generating the 3D city, it can be applied to various scenarios that require virtual 3D environments. As shown in Figure 6, we first trim incomplete buildings at the edges to ensure structural integrity. We then remove incomplete buildings at the edges and merge the remaining ones to form a larger urban environment. To further enhance the scene, we refine it using Blender by applying lighting and adding a background to enhance realism. As shown in the figure, the generated city exhibits high visual fidelity and rich texture details, making it suitable for applications such as game development, autonomous driving simulation, and mapping services. This significantly reduces the time and effort required for manual 3D modeling.

Figure 6. **Application scenarios.** After cropping and combining the generated city blocks, we enhance them in Blender by adding background and lighting. This process significantly reduces the time required by modelers to build urban environments. The high-quality results can be used in various fields, such as gaming, urban simulation, and more.

8. Additional qualitative results

Figure 7 showcases our additional qualitative results, demonstrating our method's capability in generating high-quality 3D cities with diverse stylistic variations. From a geometric perspective, our method generates buildings with varied architectural structures that conform to real-world urban designs, such as diverse rooftop designs, window arrangements on building surfaces, and vehicles on the

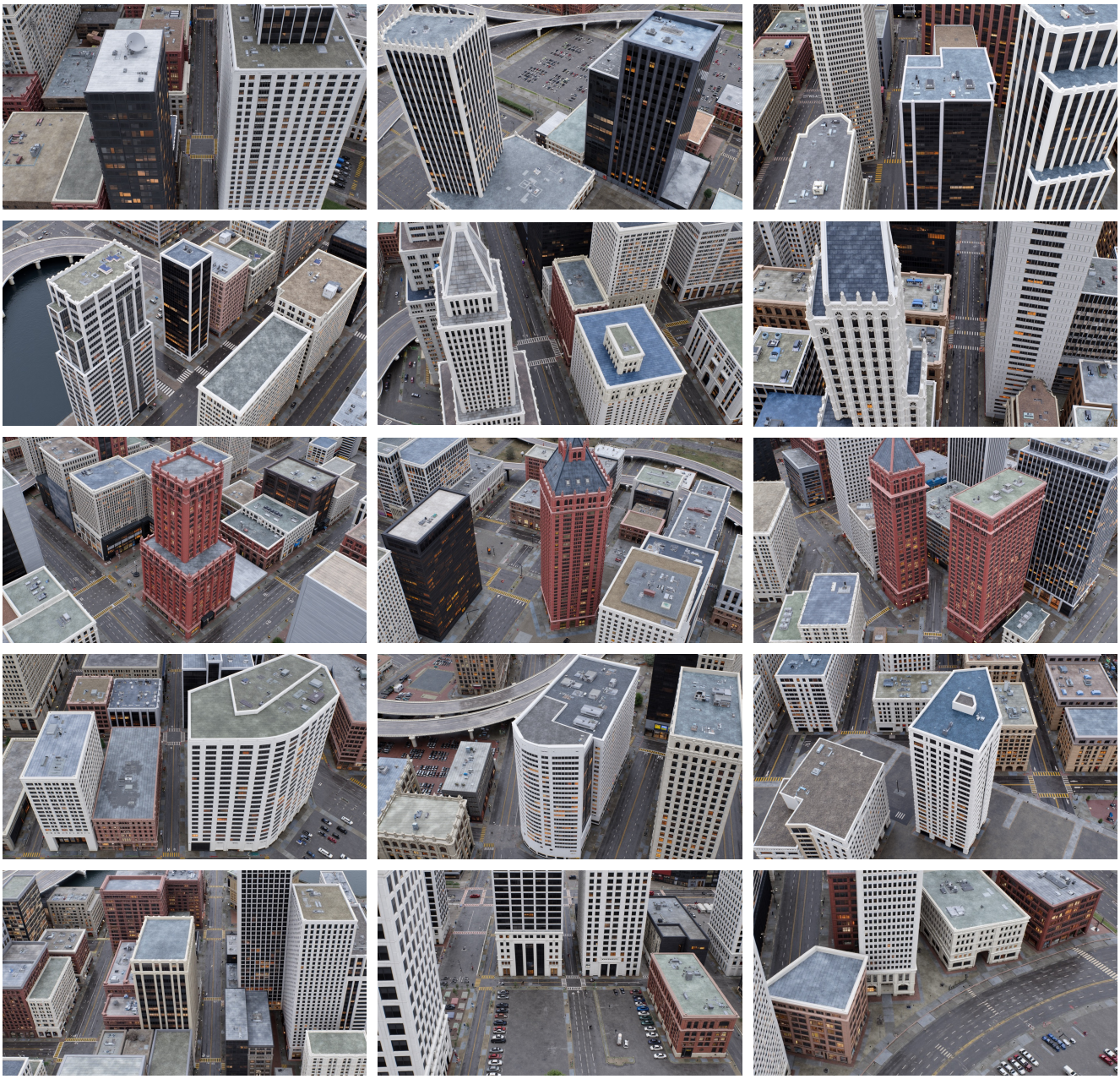

Figure 7. **Additional qualitative results.** Our method generates photorealistic cities with diverse styles. These images are generated from our city generation model.

ground. In terms of texture, our generated buildings contain rich details. As shown in the figure, our generated windows, road lane markings, and vehicles all exhibit realistic effects. Furthermore, the urban layout also aligns with real-world configurations because we generate it using satellite imagery as constraints, eliminating the need for designers to manually create layouts. These results confirm our approach's effectiveness in capturing both the geometric complexity and textural richness required for photorealistic 3D urban environment synthesis.