Satellite to GroundScape - Large-scale Consistent Ground View Generation from Satellite Views

Supplementary Material

In this supplementary material, we present further details on the methodology (Sec. 1), provide comprehensive information about the dataset (Sec. 2), report additional experimental results (Sec. 3), and discuss the limitations of our approach (Sec. 4).

1. Additional methodology details

1.1. Scene initialization and representation

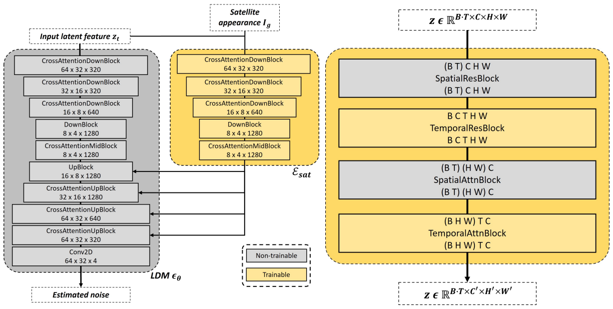
Our approach initializes the 3D scene in a format that prioritizes preserving the fidelity of the original satellite data, enabling effective camera control and ground-view generation. Unlike existing methods such as InfinitCity [5] and Sat2Scene [4], which utilize sparse point clouds, we represent the scene as a unified triangle mesh $\boldsymbol{M} = (\boldsymbol{V}, \boldsymbol{F}, \boldsymbol{E}, \boldsymbol{F}_{UV})$. This mesh consists of vertices (3D positions), faces, edges, and texture coordinates that define the appearance of each face, providing a denser and more comprehensive representation of the scene's geometry, appearance, and visibility compared to sparse point clouds. We begin with a collection of satellite images $\{\boldsymbol{I}^i\}$ and use traditional multi-view stereo (MVS) algorithms [2] to generate depth maps $\{\boldsymbol{D}^i\}$. These depth maps are projected and fused into 3D space to form an initial point cloud, $\boldsymbol{P}_0$. Using the geometry refinement method described in [10], $\boldsymbol{P}_0$ is refined to $\boldsymbol{P}_1$, which better captures building facades. The refined point cloud $\boldsymbol{P}_1$ is then triangulated to produce the triangle mesh $\boldsymbol{M} = (\boldsymbol{V}, \boldsymbol{F}, \boldsymbol{E})$. Finally, texture coordinates $\boldsymbol{F}_{UV}$ are computed by mapping the satellite images $\{\boldsymbol{I}^i\}$ onto $\boldsymbol{M}$ using texture mapping techniques [6].

1.2. Ground view foundation model

Our framework is built upon a pre-trained LDM [7], specifically leveraging a UNet-based architecture, ϵ_θ, trained on the large-scale text-image dataset LAION-5B [8]. This model has demonstrated robust capabilities in generating high-fidelity images within the ground-view domain [7]. Our objective is to guide ϵ_θ effectively to synthesize ground-view images with layouts and appearances that accurately correspond to the input satellite views and their associated camera poses. To achieve this, we use ϵ_θ as a foundational model and integrate additional modules to bridge the domain gap between satellite and ground-view imagery.

1.3. Network architecture

We show the network architecture of satellite-guided denoising and satellite-temporal denoising process, ϵ_θ, $\mathcal{E}_{sat}$ and ϵ_ϕ in Fig. 1. The LDM ϵ_θ employs a standard UNet architecture. The $\mathcal{E}_{sat}$ is designed to encode the satellite appearance $\boldsymbol{I}_g$ using a combination of "CrossAttentionDownBlock", "DownBlock" and "CrossAttentionMidBlock". For satellite-temporal denoising module ϵ_ϕ, we enhance the existing UNet structure by incorporating "TemporalResBlock" and "TemporalAttnBlock". The "TemporalResBlock" contains several 3D convolution layers, while the "TemporalAttnBlock" performs attention operations across the temporal axis (T) to achieve inter-frame learning.


Input latent feature z_t
Satellite appearance I_g
CrossAttentionDownBlock 64 x 32 x 320
CrossAttentionDownBlock 32 x 16 x 320
CrossAttentionDownBlock 16 x 8 x 640
DownBlock 8 x 4 x 1280
CrossAttentionMidBlock 8 x 4 x 1280
UpBlock 16 x 8 x 1280
CrossAttentionUpBlock 32 x 16 x 1280
CrossAttentionUpBlock 64 x 32 x 640
CrossAttentionUpBlock 64 x 32 x 320
Conv2D 64 x 32 x 4
LDM ϵ_θ
Estimated noise
CrossAttentionDownBlock 64 x 32 x 320
CrossAttentionDownBlock 32 x 16 x 320
CrossAttentionDownBlock 16 x 8 x 640
DownBlock 8 x 4 x 1280
CrossAttentionMidBlock 8 x 4 x 1280
$\mathcal{E}_{sat}$
Non-trainable
Trainable
$z \in \mathbb{R}^{B\cdot T\times C\times H\times W}$
(B T) C H W
SpatialResBlock
(B T) C H W
B C T H W
TemporalResBlock
B C T H W
(B T) (H W) C
SpatialAttnBlock
(B T) (H W) C
(B H W) T C
TemporalAttnBlock
(B H W) T C
$z \in \mathbb{R}^{B\cdot T\times C'\times H'\times W'}$

(a) Architecture of ϵ_θ and $\mathcal{E}_{sat}$

(b) Architecture of a block of ϵ_ϕ

Figure 1. **Network architecture of** ϵ_θ**,** $\mathcal{E}_{sat}$ and ϵ_ϕ. We provide detailed network architecture as the additional details to Fig. 3 & Fig. 4 in the main manuscript.

1.4. Training & inference

The training process consists of two stages: in the first stage, $\mathcal{E}_{sat}$ is trained using paired satellite appearance and ground truth images over 50,000 iterations with a batch size of 32. This stage enables $\mathcal{E}_{sat}$ to extract high-level scene layout information to guide the LDM. In the second stage, $\mathcal{E}_\phi$ and ϵ_ϕ are trained on paired sequences of five satellite appearance views and their corresponding ground truth over 100,000 iterations with a batch size of 12. This stage allows $\mathcal{E}_\phi$ and ϵ_ϕ to learn the motion and temporal features. The inference process involves satellite-conditioned and temporal-conditioned denoising, as detailed in Sec. 3.3. We adopt the deterministic sampling process from DDIM [9] with 20 denoising steps under the classifier-free guidance framework [3].

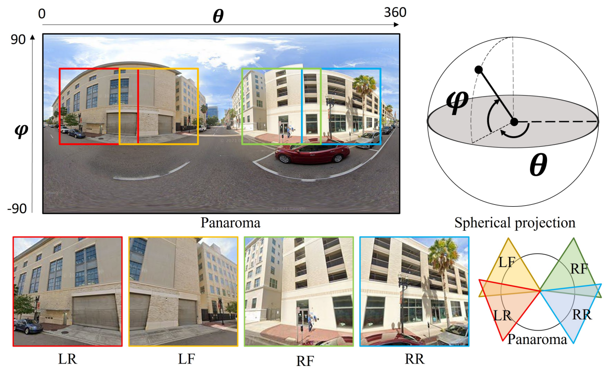


Figure 2. **Panaroma to perspective views.** The raw ground data is provided in panoramic format, where the coordinates of each pixel are defined by θ and ϕ in spherical projection. Each perspective view is resampled from a region of the spherical surface within $[\theta_{min}, \theta_{max}]$, $[\phi_{min}, \phi_{max}]$.

2. Additional dataset details

2.1. Ground data processing

The raw ground data is obtained via the GoogleStreetView API, which provides panoramic images at a resolution of 1024x2048, along with 3D location (longitude, latitude, elevation), and horizontal orientation (heading angle). The transformation process for converting each panoramic image into perspective views is detailed using camera intrinsic and extrinsic parameters. We define each perspective view within a panorama using the parameters θ, ϕ, FOV, H, W. Here, $\theta \in [0, 360]$, and $\phi \in [-90, 90]$ specify the orientation of the perspective image in azimuth and altitude angles, respectively. FOV denotes the field of view, which implicitly determines the focal length, H, and W represent the height and width of the perspective image. Specifically, we resample the "LR," "LF," "RF," and "RR" perspective images for each panorama using the parameters $\theta = [60, 120, 240, 300]$, $\phi = 15$, $FOV = 75, H = 256, W = 256$.

2.2. Train & test data split

The dataset spans a total area of $130KM^2$, divided into 90 tiles, each covering a 600×600 region, as illustrated in Fig. 3. For fair evaluation, the tiles are partitioned into training and testing sets in a 70/20 ratio.

3. Additional experiment results

Additional qualitative baseline comparisons are presented in Fig. 5, serving as an extension of Fig. 6 in the main manuscript. The results demonstrate that our method consistently generates photorealistic ground views that maintain coherence across neighboring perspectives.

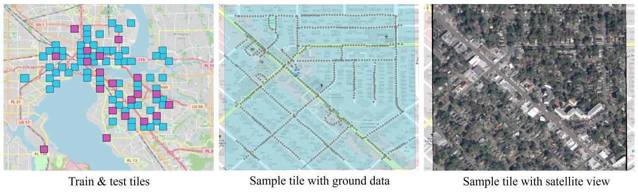

Train & test tiles
Sample tile with ground data
Sample tile with satellite view

Figure 3. **Train & test data split.** The data is split into training and testing sets based on tiles, with each tile covering a 600 × 600 m area. A total of 90 tiles are used, with 70 tiles designated for training (in blue) and 20 tiles for testing (in pink).

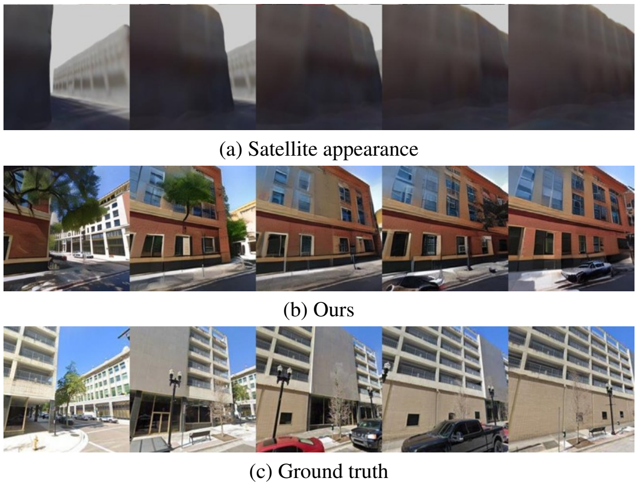

(a) Satellite appearance
(b) Ours
(c) Ground truth

Figure 4. **Limitation**. In shadow regions, our satellite-guided denoising process struggles to extract meaningful texture features for the initial ground view generation, leading to incorrect textures and facade layouts. Furthermore, the erroneous initial ground view affects the generation of subsequent views in the sequence.

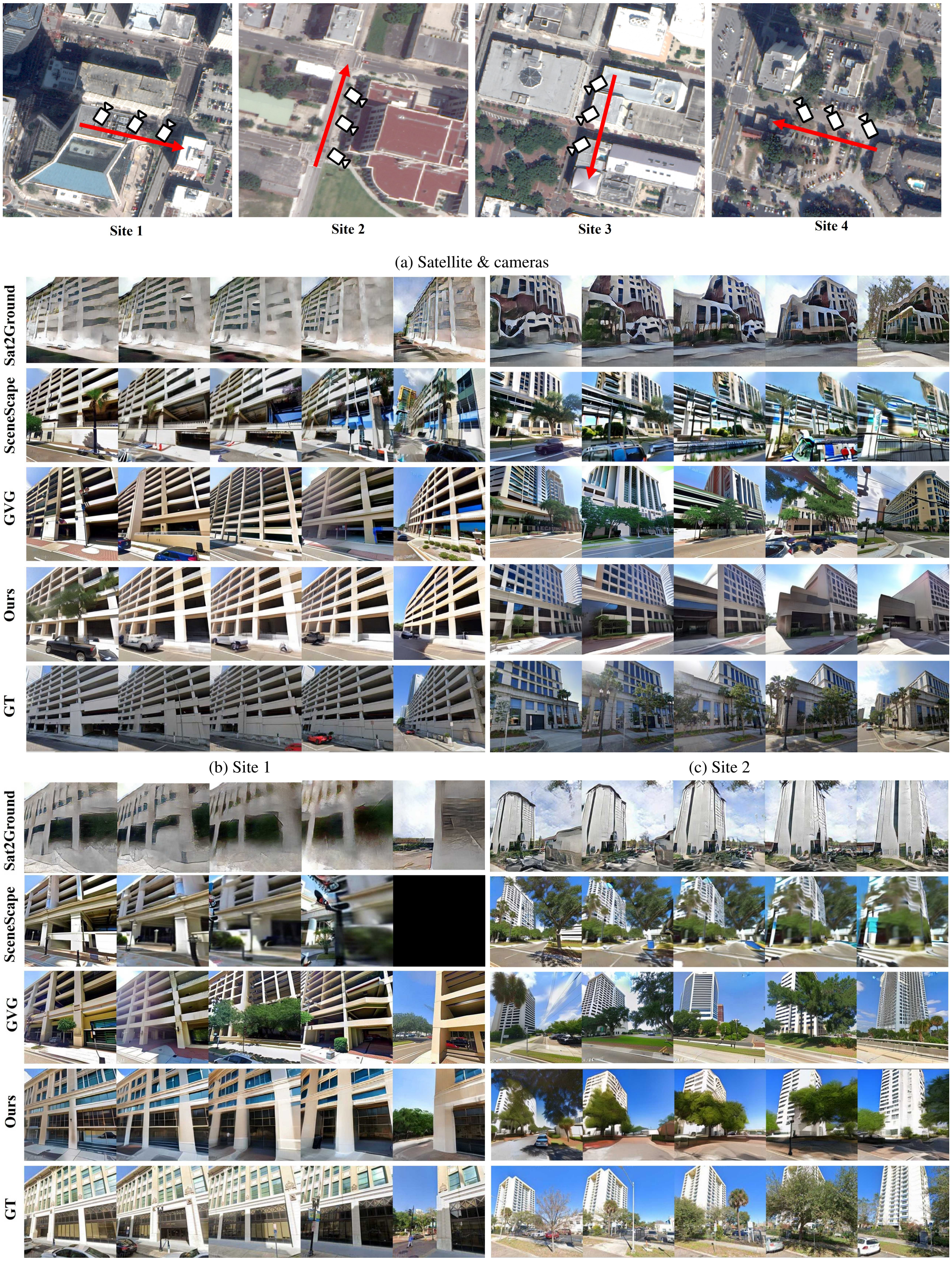
4. Limitation

Although Sat2GroundScape generates photorealistic ground views with strong multi-view consistency, it has certain limitations, as illustrated in Fig. 4. When satellite views provide insufficient appearance information, such as in shadowed or textureless regions, the $\mathcal{E}_{sat}$ module in the satellite-guided denoising process is unable to extract useful features for ground view generation, resulting in random ground layouts or appearances. This limitation also affects the satellite-temporal denoising process, leading to inconsistencies across the entire sequence of ground views.

References

[1] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: text-driven consistent scene generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 3

[2] Heiko Hirschmuller. Accurate and efficient stereo processing



(a) Satellite & cameras

(b) Site 1

(c) Site 2

(d) Site 3

(e) Site 4

Figure 5. **Additional qualitative baseline comparison on the Sat2GroundScape dataset.** Our method generates more photorealistic and consistent ground views compared to the baseline methods. Notably, SceneScape [1] in site 3 produces a black view due to the absence of warped content from the previous view.

by semi-global matching and mutual information. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 807–814. IEEE, 2005. 1

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[4] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R. Oswald. Sat2scene: 3d urban scene generation from satellite images with diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7141–7150, 2024. 1

[5] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. Infinicity: Infinite-scale city synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22808–22818, 2023. 1

[6] Xiao Ling and Rongjun Qin. Large-scale and efficient texture mapping algorithm via loopy belief propagation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[10] Ningli Xu and Rongjun Qin. Geospecific view generation–geometry-context aware high-resolution ground view inference from satellite views. *arXiv preprint arXiv:2407.08061*, 2024. 1