

Skyeyes: Ground Roaming using Aerial View Images

Zhiyuan Gao^{1,2,*} Wenbin Teng^{1,2,*} Gonglin Chen^{1,2} Jinsen Wu^{1,2}
 Ningli Xu³ Rongjun Qin³ Andrew Feng² Yajie Zhao^{1,2,†}

¹University of Southern California ²Institute for Creative Technologies ³The Ohio State University
 {gaozhiyu, wenbinte, gonglinc, jinsenwu}@usc.edu
 {xu.3961}@buckeyemail.osu.edu {Qin.324}@osu.edu {feng, zhao}@ict.usc.edu

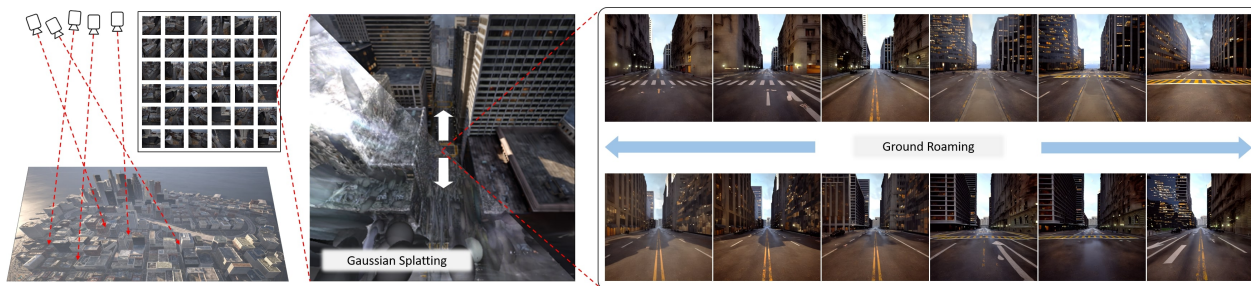


Figure 1. We proposed SkyEyes, a novel framework for efficient aerial-to-ground cross-view synthesis, transforming aerial imagery into realistic street view image sequence. This first-of-its-kind method for large-scale outdoor scenes combines 3D Gaussian Splatting with diffusion models to identify data gaps. Our constrained optimization strategy and View Consistent Module enable us to achieve images from entirely different perspectives compared to the input imagery, significantly enhancing the quality of ground-level view synthesis.

Abstract

Integrating aerial imagery-based scene generation into applications like autonomous driving and gaming enhances realism in 3D environments, but challenges remain in creating detailed content for occluded areas and ensuring real-time, consistent rendering. In this paper, we introduce Skyeyes, a novel framework that can generate photorealistic sequences of ground view images using only aerial view inputs, thereby creating a ground roaming experience. More specifically, we combine a 3D representation with a view consistent generation model, which ensures coherence between generated images. This method allows for the creation of geometrically consistent ground view images, even with large view gaps. The images maintain improved spatial-temporal coherence and realism, enhancing scene comprehension and visualization from aerial perspectives. To the best of our knowledge, there are no publicly available datasets that contain pairwise geo-aligned aerial and ground view imagery. Therefore, we build a large, synthetic, and geo-aligned dataset using Unreal Engine. Both qual-

itative and quantitative analyses on this synthetic dataset display superior results compared to other leading synthesis approaches. See the project page for more results: chaoren2357.github.io/website-skyeyes/.

1. Introduction

Creating large-scale, high-quality 3D simulation environments is crucial for applications like autonomous driving, gaming, and robotics. However, traditional methods in the gaming industry often rely on labor-intensive handcrafting, which is both time-consuming and costly, limiting their scalability and realism in depicting real-world landscapes.

Aerial imagery plays a significant role in addressing this challenge due to its wide coverage and ease of acquisition. It provides a practical resource for generating large-scale 3D terrains and environments. However, transforming aerial views into accurate ground-level views remains a complex problem due to the significant differences between aerial and ground perspectives.

Existing techniques in related areas, while effective in some contexts, face significant limitations when applied to our task. First, methods like Structure from Motion

*Equal Contribution

†Corresponding Author

(SfM) [28], Neural Radiance Fields (NeRF) [21, 31, 33, 38, 39], and 3D Gaussian Splatting (3DGS) [8, 14, 16, 19, 25] are designed for 3D reconstruction and novel view synthesis. These techniques work well when both the input and output belong to the same domain, such as generating novel ground-level views from multiple ground-level images. However, the aerial view captures the tops of buildings and large-scale layouts, revealing patterns and structures invisible from the ground, while the ground view focuses on building facades, entrances, and details like storefronts that are hidden from above. Since our task involves generating ground-level views from aerial images, which are in a different domain, these methods struggle to maintain high-quality outputs.

Second, satellite-to-ground inference techniques use satellite imagery to generate ground-level views [4, 15, 17, 18, 20, 23, 24, 32, 40]. While these methods can maintain geometric consistency between the high-altitude satellite images and the inferred ground views, they are not required to capture the same level of geometric and textural detail that our task demands. The relatively high altitude of satellite images makes these techniques insufficient for generating precise and realistic ground-level views from lower-altitude aerial inputs.

Lastly, control-based image/video generation methods [2, 3, 11, 27, 42] use aerial views to guide the generation of corresponding ground-level images. While these approaches can generate ground views that align with individual aerial images, they often struggle to maintain geometric continuity across sequences. Even if they ensure consistency between a single aerial image and its corresponding ground view, they fail to preserve coherence when generating entire sequences of ground-level views from aerial image sequences.

To address the challenges outlined earlier, we introduce Skyeyes, a framework designed to generate photo-realistic and content-consistent ground-level image sequences from aerial image inputs. As depicted in Figure 2, our approach first utilizes SuGaR, capitalizing superior detail retention ease of incremental updates, as well as its surface alignment nature compared with traditional 3DGS. This method effectively processes aerial view images and corresponding camera poses to train the model. Consequently, the optimized 3D Gaussians are then rendered from ground view perspectives, synthesizing ground view images that, while noisy, are imbued with a 3D-aware quality. Next, we implement an appearance control module designed to address the issue of preserving pixel accuracy in aerial views, a challenge noted in our integration of generative models. This module, functioning similarly to ControlNet [42] in the U-Net of the Stable Diffusion model, allows for controllable generation of photorealistic street view images. It effectively overcomes the limitations of pixel preservation in aerial imagery, en-

sureing a higher fidelity in the generated images. Finally, we introduce a view consistency module, which incorporates the concept of temporal modeling [3, 12] into the appearance control module. This integration ensures that the content generated from each view in the ground sequence maintains spatial and temporal consistency. This approach directly addresses the challenge of maintaining a consistent view within a single scene, as highlighted in the discussion of generative models. By integrating these modules, we ensure that our terrain models not only capture the intricate details of the terrain but also maintain coherence and continuity across different views.

To the best of our knowledge, there are currently no publicly available datasets that provide pairwise geo-aligned aerial and ground level image sequences. However, such dataset is crucial for training our model. To tackle the problem of data scarcity, we extract large synthetic training data from open-source simulators including CARLA [5] and CitySample [6] project developed in Unreal Engine 5. We customize sequence trajectories with respect to the location of different streets and render the scene with a spawnable camera. We will discuss more details of the dataset collection in Section 4.1. We carried out extensive experiment on the extracted datasets to compare with the traditional methods and conduct ablation studies on different components of our proposed pipeline. Results and more details will be discussed in Section 4.4 and Section 4.5. Both qualitative and quantitative analysis demonstrate that our method is superior than the other state-of-the-art frameworks. Code and both datasets will be released upon paper acceptance.

2. Related Works

2.1. Aerial-to-ground View Synthesis

Previous research primarily employs GANs [7] for generating domain-invariant features. Regmi et al. [24] and Deng et al. [4] use conditional GANs to learn ground level RGB images but overlook geometric transformations, leading to distorted outputs. [18, 20] generate panoramic depth and semantic maps using a geo-transformation layer. Toker et al. [32] apply polar transformation for satellite to ground view conversion. However, these methods often produce panoramic images with low resolution and limited details. Jang et al. [15] design a semantic-attentive transformation module for aerial to ground alignment, but focus mainly on rural areas with fewer semantic classes, while urban areas present more complex challenges. Although methods like [18, 20] produce photo-realistic street view images, they refrain from the controllable generation of textures guided by aerial view priors. Therefore, their problem settings are different from ours. To incorporate more prior knowledge of satellite imagery, Xu et al. [40] incorporate both texture and high-frequency layout as condition of the ground view

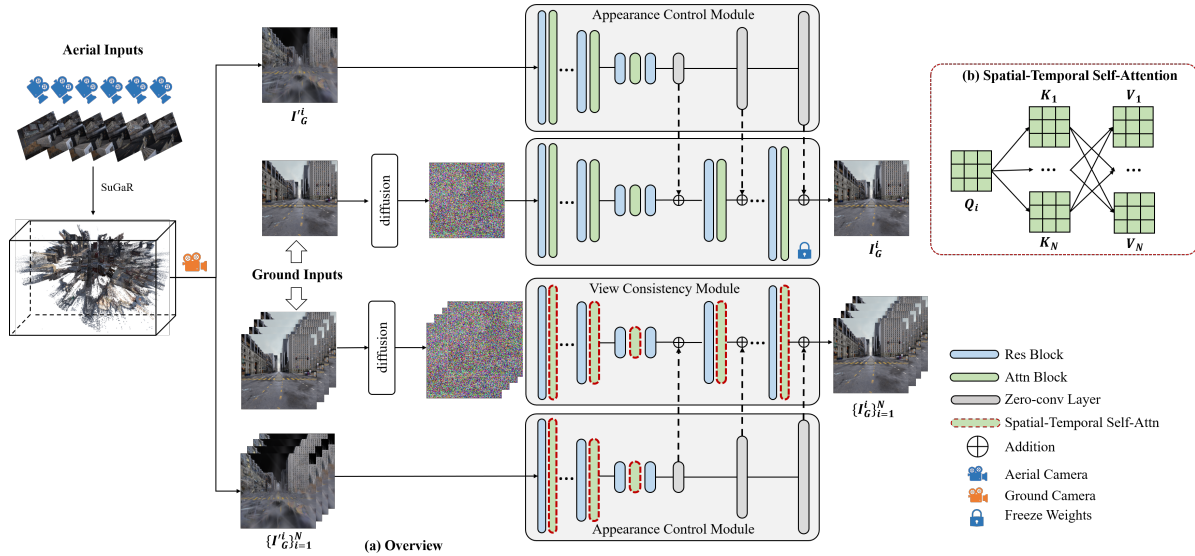


Figure 2. **(a) Overview of Skyeyes Pipeline:** Our approach commences with the utilization of SuGaR [8]. This stage involves processing aerial images and camera poses to train the model for generating ground view priors. After that, we train an appearance control module to generate photo-realistic street images **(b) Spatial-Temporal Self-Attention Module:** In the final stage, our view consistency module integrates temporal modeling to ensure spatial and temporal coherence across different views. This module, akin to a spatial-temporal self-attention mechanism, guarantees the consistency and continuity of the scene’s depiction across various perspectives. At inference time, given a sequence of ground view priors rendered from SuGaR [8], our view consistency module can generate photo-realistic and temporal consistent ground view sequence by denoising from pure Gaussian noise.

panorama generation model, but it fails to address the time consistency problem across different frames.

2.2. Large Scale Novel View Synthesis

Novel view synthesis, primarily driven by Neural Radiance Fields (NeRF) [21], has seen significant advancements through deep learning, enabling diverse scene representations and new view rendering. However, traditional NeRF struggles with large-scale environments due to intense memory and computational demands and challenges in handling transient objects. To address these limitations, recent developments have focused on adapting NeRF techniques for ground-level and aerial-level perspectives. From ground-level perspective, Block-NeRF [31] subdivides large environments into smaller, independently trained NeRFs, but it still encounters issues like temporal inconsistencies and less detailed reconstructions of distance objects. Scalable Urban Dynamic Scenes(SUDS) [34] offers a novel approach by factorizing scenes into static, dynamic, and far-field components using separate hash tables to solve challenges in dynamic elements. But the work only focuses on urban settings. StreetSurf [9], on the other hand, separates unbounded spaces into multi-view segments and utilizes hyper-cuboid hash-grids and a road surface initialization scheme to enhance representation. However, this method is primarily tailored for autonomous driving datasets and may under-perform in poor lighting conditions.

Street Gaussians [41] offers a different approach for urban scenes using 3D Gaussians, enabling swift object and background composition, but are limited to grid dynamics. Despite all these innovations, handling transient objects still remains as a challenge in the field.

In the aerial perspective domain, several studies have extended NeRF-based methodology to encompass larger scenes. Xiangli et al. [38] employed a progressive training strategy for NeRF models to handle multi-scale scenes. Xu et al. [39] developed a two-branch architecture with a feature grid for efficient rendering in large city scenes. Turki et al. [33] proposed a geometric clustering method for parallel training of NeRF submodules. However, these NeRF-based approaches struggle with realistic image generation from significantly different viewpoints due to the limited range of input perspectives.

2.3. Controllable Image and Video Diffusion Model

Diffusion model and latent diffusion model has exhibited their effectiveness in conditional image generation. By simply adding a text prompt, methods like Imagen [27] and Stable Diffusion (SD) can achieve the ideal customization of content synthesis. The controllable image generation has been largely extended with the advent of ControlNet [42], which allows additional condition to SD models such as depth, pose and segmentation maps. Established on ControlNet, Control-A-Video (CAV) [3] generates both con-

trollable and content-consistent video based on sequence of control maps and text conditions. Apart from the traditional 3D U-Net proposed by video diffusion model [11], one of the main contributions of CAV is the introduction of motion-adaptive noise initializer, which preserves the latent similarity between frames as appose to the random Gaussian noise.

3. Skyeyes

In this section, we elaborate on the details of Skyeyes. Given a sequence of aerial imagery $\{I_A^i\}_{i=1}^N$ and corresponding camera pose $\{W_A^i\}_{i=1}^N$, our goal is to synthesize a sequence of ground image $\{I_G^i\}_{i=1}^N$ conditioned on ground camera poses $\{W_G^i\}_{i=1}^N$, where N is the number of selected frames in a sequence. The synthesized images $\{I_G^i\}_{i=1}^N$ should maintain its content coherence.

The overall architecture of Skyeyes is shown in Figure 2. We will first introduce the preliminaries of our proposed method in Section 3.1, which includes 3D Gaussian Splatting and latent diffusion model/ControlNet. Then we will introduce our method in two steps. The first step, which involves the Appearance Control Module, is presented in Section 3.2. The second step, concerning the View Consistency Module, is detailed in Section 3.3.

3.1. Preliminary

3.1.1 Surface-Aligned 3D Gaussian Splatting(SuGaR)

3DGS [16] models a scene as a set of differentiable 3D Gaussians that could be easily rendered with tile-based rasterization. Each Gaussian is parameterized by a center point μ_g and a covariance matrix Σ_g :

$$G(x) = e^{-\frac{1}{2}(x-\mu_g)^T \Sigma_g^{-1}(x-\mu_g)} \quad (1)$$

When rendering, the color and opacity of all the Gaussians are calculated by Equation 1. The final pixel color C is computed by blending all the 2D Gaussians that contributes to the pixel:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (2)$$

where c_i and α_i are the view dependent color and opacity of the Gaussian. For more details, we recommend the original work from [16].

In the SuGaR framework that utilizes 3D Gaussian Splatting, the process begins by incorporating a loss term based on the signed-distance field (SDF). This loss term, represented by Equation 3, ensures the alignment of 3D Gaussians with the scene’s surface during optimization. A rough mesh is extracted from the aligned Gaussians, and both the mesh and the 3D Gaussians situated on the mesh surface are optimized jointly using Gaussian Splatting rendering, resulting in a new set of Gaussians that are tied to an editable mesh.

$$R = \frac{1}{|P|} \sum_{p \in P} \left| \frac{\langle p - \mu_{g^*}, n_{g^*} \rangle}{s_{g^*}} - \frac{\langle p - \mu_g, n_g \rangle}{s_g} \right| \quad (3)$$

Here, R denotes the residual error across a set of sample points P . s_g is the smallest scaling factor of Gaussian g , which signifies how flat the Gaussian is—approaching zero implies increased flatness. The parameters μ_{g^*} and s_{g^*} are the optimal Gaussian parameters that align best with the scene’s surface.

Moreover, the methodology seeks to avoid semi-transparent Gaussians to accurately describe the scene’s surface, hence, the opacity coefficient α_g is set to 1 for any Gaussian g .

3.1.2 Latent Diffusion Models and ControlNet

Compared to diffusion models [11, 30], latent diffusion models [26] synthesize features of images in a latent space defined by a pre-trained autoencoder. A common schema is to add textual inputs into image generation by converting a text prompt into embeddings c_{text} . This is usually achieved by a CLIP-based transformer for text encoding. Given an Image I and encoder \mathcal{E} , the initial latent feature $z_0 = \mathcal{E}(I)$ is perturbed by a sequence of Gaussian noise such that after T steps the latent feature z_T fall in a standard Gaussian distribution $\mathcal{N}(0, 1)$. The objective of latent diffusion model is to optimize a denoising process formulated by a U-Net architecture:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0, c_{text}, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, c_{text}, t)\|_2^2 \right] \quad (4)$$

Here $t = 1 \dots T$ is the time embedding during denoising process. ϵ is a standard normal distribution and ϵ_θ is the neural network parameterized by θ .

ControlNet [42] further boost the controllability of latent diffusion model by adding a specific image condition such as depth or semantic map. The downsampling blocks and middle block of ControlNet is a trainable copy of Stable Diffusion [26] whereas its main contribution is to add a series of zero-convolutions whose outputs are added to the skipped connection of Stable Diffusion U-Nets. Suppose the task-specific image condition is denoted as c_f , the objective is formulated as follows:

$$\mathcal{L}_{Control} = \mathbb{E}_{z_0, c_{text}, c_f, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, c_{text}, c_f, t)\|_2^2 \right] \quad (5)$$

3.2. Appearance Control Module

The objective of appearance control module is to leverage the controllable image generation ability of LDM [26] to generate $\{I_G^i\}_{i=1}^N$ conditioned on $\{I_A^i\}_{i=1}^N$. This naturally leads us to the architecture of ControlNet [42]. However, ControlNet incorporates task-specific images (depth, semantic, etc.) as conditions which are not accessible for

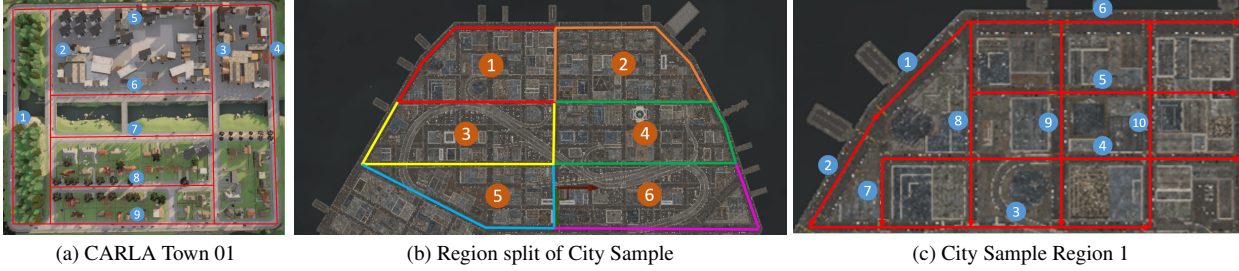


Figure 3. Comprehensive visual representation of the data extraction process from CARLA and City Sample Project.

the generation of ground view images. One straightforward way is to project pixels of I_A^i from image space to world space and project back to the image space of $\{I_G^i\}$, but this requires accurate geometry of I_A^i , which is usually hard to obtain in real application and a coarse depth estimation will exhibit large view distortion and appearance distinction.

In appearance control module, we propose to leverage SuGaR [8] to construct the scene given both $\{I_A^i\}_{i=1}^N$ and $\{W_A^i\}_{i=1}^N$. Compared with traditional 3DGS [16] tends to align the 3D Gaussians with the surface of the object. The optimized 3D Gaussians are rendered with ground view cameras $\{W_G^i\}_{i=1}^N$ to synthesize ground view control maps $\{I_G^i\}$:

$$I_G^i = \mathcal{R}(G(\{I_A^i\}_{i=1}^N, \{W_A^i\}_{i=1}^N), W_G^i) \quad (6)$$

where \mathcal{R} is the Gaussian splatting renders. In this step, we first initialize the point cloud using Structure-from-Motion (SfM). The inputs for SfM are the sets $\{I_A^i\}_{i=1}^N$ and $\{W_A^i\}_{i=1}^N$, as accurate camera poses are essential for geo-registration.

3.3. View Consistency Module

With the appearance control module discussed in Section 3.2 we are able to synthesize a photo-realistic ground view image from a noisy 3D Gaussian prior. Nevertheless, how to ensure content consistency across all the views in a sequence remains a challenging problem. The general purpose of appearance control module is to refine the blurry regions and in-paint the unseen regions with the powerful content generation ability of latent diffusion model. However, how the regions are refined and in-painted may have thousands of explanations. In this work, we are inspired by vid2vid-Zero [36] and Control-A-Video [3] to propose a view consistency module (VCM) for a sequence of generated ground views. VCM is integrated to the up-sampling and down-sampling blocks of LDM to maintain both spatial and temporal consistency. Illustrated by Figure 2, suppose the self-attention calculated by the pre-trained LDM is given by:

$$\text{Self_Attn}(\cdot) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \quad (7)$$

where Q, K, V are the query, key, value features of the spatial features x such that $Q = W^Q x, K = W^K x, V =$

$W^V x$, and W^Q, W^K, W^V are the corresponding learnable projection matrix. Instead of simply considering the spatial self-attention across the feature maps, we incorporate the spatial-temporal self-attention across the frames, where the projection matrices are shared for all the frames:

$$Q = W^Q x_i, K = W^K x_{1:F}, V = W^V x_{1:F} \quad (8)$$

where x_i is the query frame, and $x_{1:F}$ are the concatenation of all the frame in a sequence, i.e. $x_{1:F} = [x_1, x_2, \dots, x_F]$. By attending both the spatial features across the feature map and the temporal features across all the frames, we find it effective to alleviate the discrepancy between each sampling process of LDM and ControlNet, with more details are discussed in Section 4.5.

With memory efficiency as well as long sequence generation purpose, we condition the generation process conditioned on the latent space of the first frame. Specifically, we add noise to each frame of the random sampled sequence except for the first frame. By this training scheme, the diffusion model can learn to generate the subsequent frames based on the first frame. The objective is formulated as follows:

$$\mathcal{L}_{\text{VCM}} = \mathbb{E}_{z_0, c_{\text{text}}, c_f, t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, c_{\text{text}}, c_f, t, z^1)\|_2^2 \right] \quad (9)$$

where z^1 is the latent feature of the first frame. At inference time, we first generate the first frame x^1 with our appearance control module and use x^1 as the condition to autoregressively generate the subsequence frames:

$$x_{2:F} = \text{VCM}(z_{2:F}, c_{\text{text}}, c_f, \mathcal{E}(x^1)) \quad (10)$$

This proposed method will refrain the diffusion model from memorizing all the frames in a video thus achieves both memory efficiency and long sequence generation targets. During training, we randomly sample a consecutive of 12 frames within a long ground view sequence and train the diffusion U-Net with the first-frame conditioning method.

4. Experiments

4.1. Dataset Collection

We utilized two distinct scene simulation platforms, CARLA Simulator [5] and CitySample [6] from Unreal

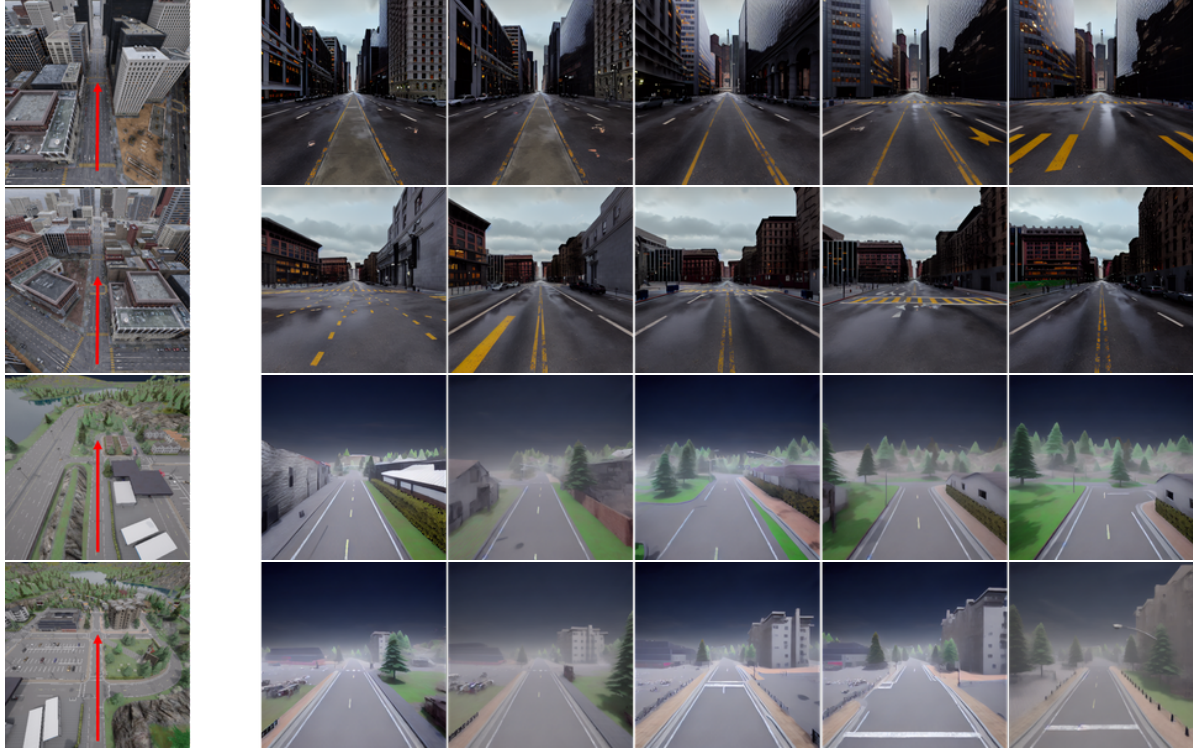


Figure 4. **Qualitative Results.** Conditioned on aerial images (leftmost column), our method synthesizes realistic and view-consistent ground view sequences. The first two rows are from the CitySample dataset, and the last two from the CARLA dataset. We strongly recommend checking the supplementary material for more results.

Engine 5, to generate geo-aligned aerial and street view datasets for detailed and complex urban or rural environments.

CARLA Simulator is an open-source platform designed for the development, training, and validation of autonomous driving systems. The sequences are extracted from various maps including Town01, Town02, Town03, Town04 and Town05 where we manually locate the start and end point of each lane (See Fig. 3a for an example of lane selection of Town 01).

CitySample project is created by Ubisoft in Unreal Engine 5. We selected the smaller city level within this project for data extraction. Given the expansive scale of the map, we segmented it into multiple regions (refer to 3b for details) and appoint multiple lanes inside a region (refer to 3c for details). Please refer to our supplementary materials for a more detailed description of data collection process.

4.2. Implementation Details

We organized aerial and ground view images by distinct lanes, treating each lane’s set as a sequential dataset with ground-level images as primary input. We first train SuGaR on an NVIDIA RTX 4090 GPU for 15K iterations, resizing images to 512×512 . Then, priors $\{I_G^i\}$ are rendered from ground view poses for all sequences. These priors are

used to train the appearance control module for 30K iterations, with a batch size of 32 on 4 NVIDIA A100 GPUs. We condition the diffusion model with a uniform text prompt, “a realistic street view image”. Finally, we freeze the appearance control module and train the view consistency module for 3K iterations with a batch size of 2, sampling 12 frames from a large sequence for generalization.

4.3. Baselines and Metrics

Baselines As mentioned in Section 1, we chose three types of baseline methods that aligned with our task. For 3D reconstruction methods, we chose MVS [28], NeRF [21], 3DGS [16] and SuGaR [8]; Geospecific View Generation (GVG) [40] for satellite-to-ground related baseline; ControlNet [42] and InstructPix2Pix [2] for control-based image/video generation baseline.

Evaluation metrics We use both image-based and video-wise metrics for understanding the efficacy of different techniques. For image-based evaluation, we considered metrics like PSNR [13], SSIM [37], LPIPS [43] and FID [10]. For video-wise metric, we use Fréchet Video Distance (FVD) [1, 35] and Kernel Video Distance (KVD) [35] to for video temporal consistency and quality evaluation.

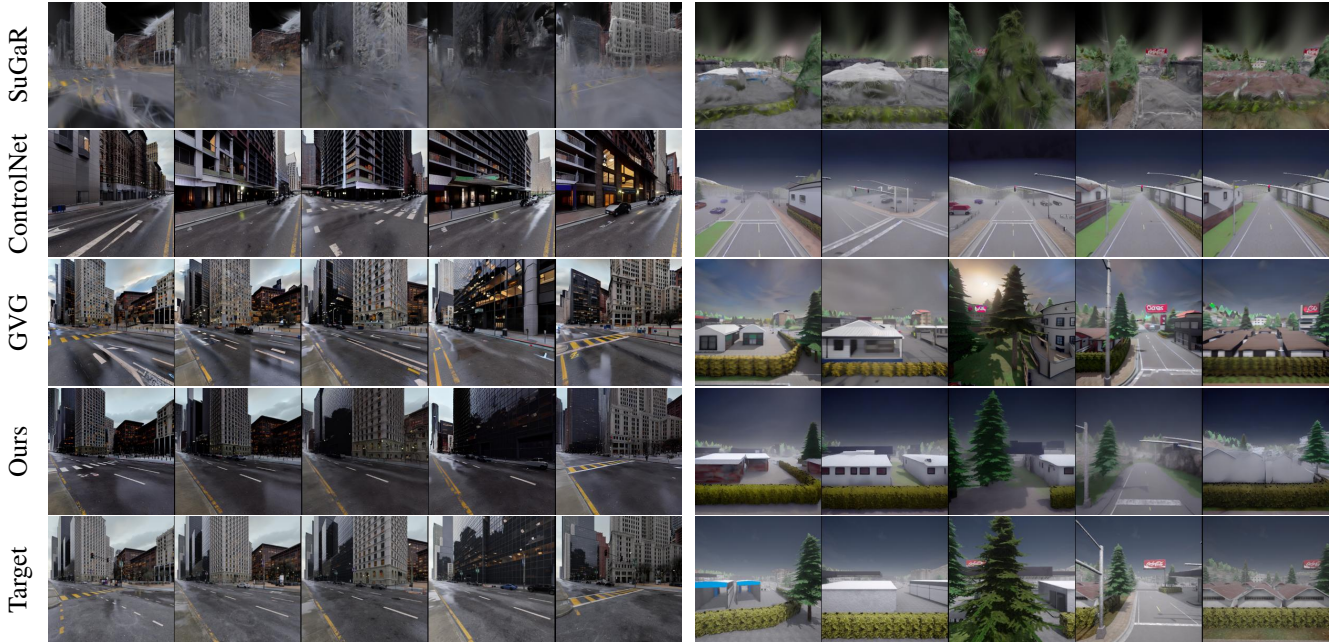


Figure 5. **Qualitative Comparisons.** We compare Skyeyes with other SOTA methods for ground view generation. Unlike tasks that require matching ground truth, our task focuses on generating visually plausible images with continuous textures. All methods were evaluated under the same conditions, and Skyeyes consistently delivers superior visual quality.

Table 1. Quantitative comparison between Skyeyes and other state-of-the-art methods on the test set of City Sample dataset.

CitySample	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	KVD ↓	FVD ↓
MVS [29]	359.15	27.79	0.30	0.63	377.20	2846.69
NeRF [22]	317.09	27.94	0.28	0.68	382.57	2390.31
3DGS [16]	245.24	28.13	0.42	0.62	340.62	1926.74
SuGaR [8]	260.51	28.13	0.38	0.60	204.20	1157.64
ControlNet [42]	63.47	28.08	0.25	0.57	281.89	1205.81
Instruct-P2P [2]	100.47	28.04	0.25	0.58	428.88	1742.12
GVG [40]	29.62	28.29	0.33	0.47	141.33	715.97
Ours	54.73	32.22	0.45	0.48	117.93	528.65

Table 2. Quantitative comparison between Skyeyes and other SOTA methods on the test set of CARLA dataset.

CARLA	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	KVD ↓	FVD ↓
MVS [29]	388.37	27.82	0.40	0.53	562.21	3606.30
NeRF [22]	248.16	27.98	0.51	0.68	618.43	2571.87
3DGS [16]	228.92	28.32	0.59	0.48	573.05	2404.44
SuGaR [8]	202.38	28.13	0.53	0.48	679.40	2498.16
ControlNet [42]	75.26	27.97	0.58	0.50	277.89	1056.69
Instruct-P2P [2]	202.12	27.80	0.38	0.65	707.08	3327.93
GVG [40]	45.73	28.29	0.53	0.47	266.46	913.07
Ours	57.95	33.37	0.69	0.44	218.29	693.28

4.4. Results

Fig. 4 visualizes the results of our proposed pipeline on two extracted datasets. Specifically, given a sequence of aerial view images, Skyeyes is able to predict and synthesize the corresponding ground view sequence.

Qualitative Comparison We present the visual performance comparison of the baselines and our method in Figure 5, showcasing the effectiveness of each approach in rendering visually realistic terrain. As observed, SuGaR, while

rendering geometry and color accurately, results in a somewhat blurred image, primarily due to splatting effects when viewed from this extreme perspective, especially in comparison to aerial views. ControlNet appears to produce photo-realistic images with less artifacts, their textures are significantly different from the ground truth images. Compared with GVG, our method produces less artifacts and largely maintain intra-frame consistency.

Quantitative Comparison A significant improvement is evident in the KVD and FVD metrics, as shown in Tables 1 and 2. Both KVD and FVD values are significantly lower in our method compared to others, as these metrics indicate better performance when the values are smaller. Our FVD, for instance, improves by around 25% on average compared to the best baseline. This substantial reduction demonstrates that our method maintains superior consistency in video sequences, ensuring smooth transitions and coherence across frames. Furthermore, in terms of image-related metrics such as FID, PSNR, SSIM, and LPIPS, our method consistently ranks first or second, showcasing its competitive edge in image generation quality. In some cases, our results significantly surpass those of other methods, further emphasizing the high-quality, realistic rendering of the terrain in both image and video aspects.

4.5. Ablation Study

View Consistency Module To demonstrate the effectiveness of the View Consistency Module, we attempted to generate results by discarding the View Consistency Module

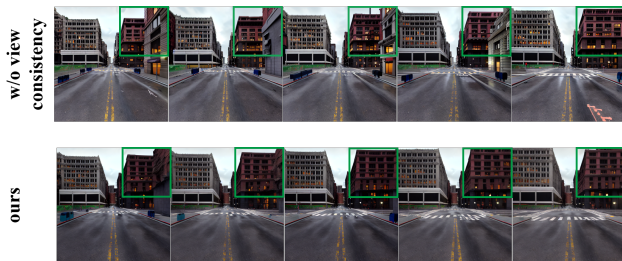


Figure 6. **Ablation Study on view consistency module.** We observe apparent content inconsistency when view consistency module is dropped (top row), whereas content remain relatively consistent for our full pipeline (bottom row). The area surrounded by green square more apparently illustrates content consistency of our full pipeline.

while using the same prior. Top row of Fig. 6 presents that if we directly use Appearance Control Module without adding the View Consistency Module, each frame produces buildings with different colors and materials. There are significant differences in the arrangement and number of windows, and even the number of tall buildings in the distance lacks continuity. With the incorporation of this module, not only is there consistency in the appearance of the buildings, but also the road markings and pedestrian crossings on the ground are somewhat consistent, greatly enhancing the continuity of the generated scene.

Different choices of ground view priors We study different model choices for ground view image prior generation. The ground view image prior is crucial in photo-realistic sequence generation as an ideal prior should possess more abundant content and features whereas a less ideal prior will present more blurry and noisy spaces. Therefore, we compare three different prior generation models: 3DGS [16], Scaffold GS [19] and SuGaR [8] (our choice for the pipeline) and evaluate all of the models on CitySample dataset. The results in Fig. 7 indicates the generation results under the same camera pose. Although the prior of vanilla 3DGS and Scaffold GS have the potential in guiding the diffusion model to generate photo-realistic images, their appearances present significant differences compared with ground truth. In comparison, the prior of SuGaR illustrates strong ability in generation of images with controllable appearance.

5. Limitations and Future Work

One primary limitations of Skyeyes is its current performance in generalizing to real-world data. The framework, as it stands, is largely trained on synthetic datasets extracted from simulators like CitySample. While these datasets offer a controlled environment for training, they may not fully capture the complexity and variability found in real-world scenarios. The textures, lighting conditions, and architec-



Figure 7. **Ablation study on prior generation model.** We compare the generation quality based on different ground view prior. Specifically, we compare ScaffoldGS [19], 3DGS [16], and SuGaR [8] (ours). The red square indicates the prior generated at the same camera pose. Though experience longer training time, the prior of SuGaR presents higher generation quality.

tural elements in synthetic environments can differ significantly from those in real-world settings. This discrepancy can lead to challenges in achieving the same level of detail and realism when the model is applied to actual aerial and ground-level imagery. Addressing this limitation will involve refining the training datasets to include more diverse and realistic scenarios.

6. Conclusion

Our research introduces Skyeyes, a groundbreaking framework designed for aerial-to-ground cross-view synthesis, adeptly transforming aerial imagery into detailed and realistic 3D terrain models. This innovative approach, a first in large-scale outdoor scene generation, skillfully integrates 3DGS with controllable diffusion models. This integration not only identifies and fills data gaps but also provides robust prior feature to the controllable generation of diffusion model. The diffusion model further enhances this framework by ensuring noise control and maintaining spatiotemporal consistency, thereby producing superior quality results compared to traditional video-to-video synthesis methods. Our experimental results demonstrate Skyeyes’ effectiveness in creating high-quality, realistic terrain models. This success is evident in the framework’s ability to surpass existing methods in terms of visual accuracy and consistency. Skyeyes stands out as a significant advancement in terrain generation.

7. Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0075. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

- [1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 6
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 6, 7
- [3] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 2, 3, 5
- [4] Xueqing Deng, Yi Zhu, and Shawn Newsam. What is it like down there? generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52, 2018. 2
- [5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2, 5
- [6] EpicGames. City sample project: Unreal engine demonstration. <https://docs.unrealengine.com/5.0/en-US/city-sample-project-unreal-engine-demonstration/>, 2023. 2, 5
- [7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2
- [8] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775*, 2023. 2, 3, 5, 6, 7, 8
- [9] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 3
- [10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2, 4
- [12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2
- [13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 6
- [14] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [15] Jinhyun Jang, Taeyong Song, and Kwanghoon Sohn. Semantic-aware network for aerial-to-ground image synthesis. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3862–3866. IEEE, 2021. 2
- [16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4, 5, 6, 7, 8
- [17] 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*, pages 7141–7150, 2024. 2
- [18] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Rongjun Qin, Marc Pollefeys, and Martin R. Oswald. Sat2vid: Street-view panoramic video synthesis from a single satellite image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12436–12445, October 2021. 2
- [19] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 2, 8
- [20] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2
- [21] 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*, 2020. 2, 3, 6
- [22] 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, July 2022. 7
- [23] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3683–3692, 2023. 2
- [24] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018. 2
- [25] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *arXiv preprint arXiv:2403.17898*, 2024. 2
- [26] 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. 4
- [27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour,

- Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3
- [28] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 6
- [29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7
- [30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4
- [31] Matthew Tancik, Vincent Casser, Xincheng Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2, 3
- [32] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 2
- [33] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12931, 2022. 2, 3
- [34] Haithem Turki, Jason Y. Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes, 2023. 3
- [35] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6
- [36] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 5
- [37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [38] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *European conference on computer vision*, pages 106–122. Springer, 2022. 2, 3
- [39] Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes, 2023. 2, 3
- [40] 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. 2, 6, 7
- [41] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. 2023. 3
- [42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 3, 4, 6, 7
- [43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6