

CROSSVIEWDIFF: A CROSS-VIEW DIFFUSION MODEL FOR SATELLITE-TO-STREET VIEW SYNTHESIS

Weijia Li^{1*} **Jun He**^{1*} **Junyan Ye**^{1,2*} **Huaping Zhong**^{2,3*}

Zhimeng Zheng² **Zilong Huang**¹ **Dahua Lin**² **Conghui He**^{2,3†}

¹ Sun Yat-Sen University, China

² Shanghai Artificial Intelligence Laboratory, China ³ SenseTime Research, China

ABSTRACT

Satellite-to-street view synthesis aims at generating a realistic street-view image from its corresponding satellite-view image. Although stable diffusion models have exhibit remarkable performance in a variety of image generation applications, their reliance on similar-view inputs to control the generated structure or texture restricts their application to the challenging cross-view synthesis task. In this work, we propose CrossViewDiff, a cross-view diffusion model for satellite-to-street view synthesis. To address the challenges posed by the large discrepancy across views, we design the satellite scene structure estimation and cross-view texture mapping modules to construct the structural and textural controls for street-view image synthesis. We further design a cross-view control guided denoising process that incorporates the above controls via an enhanced cross-view attention module. To achieve a more comprehensive evaluation of the synthesis results, we additionally design a GPT-based scoring method as a supplement to standard evaluation metrics. We also explore the effect of different data sources (e.g., text, maps, building heights, and multi-temporal satellite imagery) on this task. Results on three public cross-view datasets show that CrossViewDiff outperforms current state-of-the-art on both standard and GPT-based evaluation metrics, generating high-quality street-view panoramas with more realistic structures and textures across rural, suburban, and urban scenes. The code and models of this work will be released at <https://opendatalab.github.io/CrossViewDiff/>.

1 INTRODUCTION

Satellite images captured by high-altitude sensors differ significantly from daily images taken by ordinary ground cameras. The overhead perspective of satellite images provides a macroscopic view that encompasses extensive regional topography, building layouts, and road networks. street-view images, on the other hand, are captured by mobile phones or vehicle-mounted cameras, providing a ground-level observation of the scene. In this study, we address the task of cross-view synthesis, especially satellite-to-street view synthesis, which is an important and challenging computer vision task that has received increasing attention in recent years Shi et al. (2022); Qian et al. (2023); Lu et al. (2020). Generating realistic street-view images from corresponding satellite images through cross-view synthesis can benefit various applications, such as cross-view geolocation Li et al. (2024a); Toker et al. (2021), urban building attribute recognition Ye et al. (2024b), and 3D scene reconstruction Li et al. (2024c).

Due to the significant differences in viewpoints and imaging methods, the overlapping information between different perspectives is very limited Tang et al. (2019); Regmi & Borji (2018); Ye et al. (2024c;a), as shown in Figure 1 (a). This creates a substantial domain gap between satellite and street-view images, making the synthesis task highly challenging Lu et al. (2020); Shi et al. (2022). Consequently, some studies have explored the use of additional ground truth semantic segmentation maps as auxiliary conditions for models to improve the synthesis results Zhai et al. (2017); Regmi & Borji (2018); Tang et al. (2019); Wu et al. (2022). However, this essentially generates

*These authors contributed equally to this work.

†Corresponding author(s). E-mail(s):heconghui@pjlab.org.cn

images from semantic maps and does not truly accomplish satellite-to-street cross-modal generation. Other studies have explored various satellite-to-street projection or transformation methods, utilizing geometric structure priors derived from satellite images to enhance the layout and structure of synthesized street-view panoramas Lu et al. (2020); Toker et al. (2021); Shi et al. (2022); Qian et al. (2023). However, there has been limited exploration of the fidelity and consistency of textures in cross-modal synthesis between satellite images and street-view panoramas.

Furthermore, existing satellite-to-street view synthesis methods are mostly based on Generative Adversarial Networks (GANs), which often result in poor image quality and unrealistic textures in the synthesized results, as shown in Figure 1 (b).

Recently, diffusion models have demonstrated superior performance in various content generation applications, garnering widespread attention Song et al. (2021); Ho et al. (2020); Balaji et al. (2022); Ramesh et al. (2022); Saharia et al. (2022b). Models like ControlNet enable controllable image synthesis based on various visual conditions Zhang et al. (2023a); Huang et al. (2023b); Zhao et al. (2023); Ruiz et al. (2023). For satellite-to-street view synthesis, one potential solution is to treat this task as a controllable image synthesis task, using satellite images to control the synthesis of street-view images. However, existing methods utilize similar-view images (e.g., sketches, segmentation maps) as inputs to control the structure or texture of the generated results. The different modality domains of satellite and street-view images limit the applicability of these methods in cross-view synthesis tasks. As shown in Figure 1(b), the domain gap results in synthesized images that are often realistic yet inconsistent, with significant differences between the synthesized street-view images and the actual corresponding satellite content.

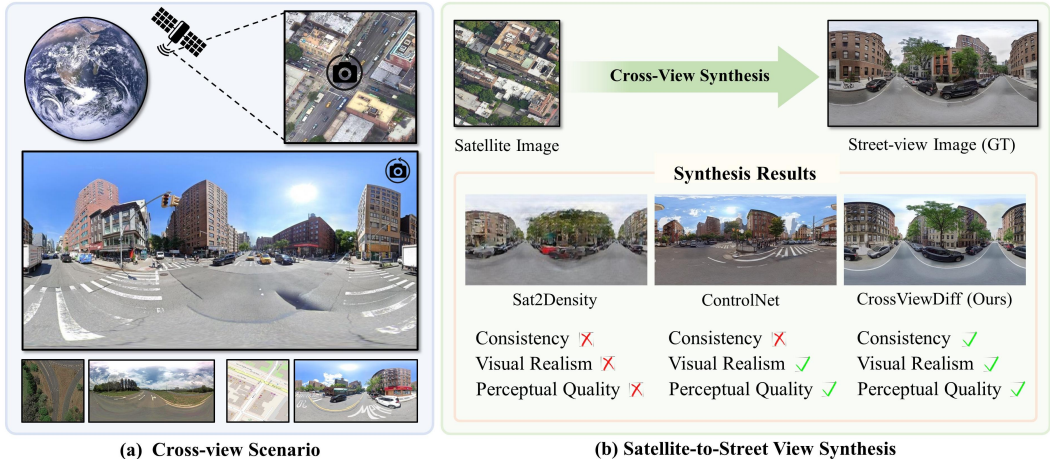


Figure 1: Illustration of the satellite-to-street view synthesis task. (a) In cross-view scenarios, the satellite view and street view differ significantly, with limited overlapping information, posing a serious challenge to the satellite-to-street view synthesis task. (b) Compared with existing methods using GANs (e.g., Sat2Density Qian et al. (2023)) or diffusion models (e.g., ControlNet Zhang et al. (2023a)), CrossViewDiff is capable of synthesizing more realistic street-view images with better perceptual quality and consistency with Ground Truth.

Furthermore, existing cross-view generation studies Lu et al. (2020); Toker et al. (2021); Regmi & Borji (2018) commonly use image generation metrics such as SSIM Wang et al. (2004) and PSNR to evaluate the content consistency of synthesized images, as well as FID Heusel et al. (2017) and KID Bińkowski et al. (2018) to assess image realism. However, these traditional metrics often fall short in aligning with human perception and lack transparency and interpretability. With the development of multimodal large models (MLLM) OpenAI (2023); Team et al. (2023); Liu et al. (2024); Li et al. (2023b), an increasing number of studies have employed multimodal large models like GPT-4o OpenAI (2023) for assessing the quality of synthesized images Cho et al. (2023); Huang et al. (2023a); Wu et al. (2024); Zhang et al. (2023b), achieving interpretable and highly human-aligned scoring Ku et al. (2023); Peng et al. (2024). However, prior use of multimodal scoring

has predominantly been in text-to-image synthesis or editing tasks, with no studies applying it to cross-view synthesis tasks.

In this work, we propose CrossViewDiff, a cross-view diffusion model for satellite-to-street view synthesis. Based on the geometric and imaging relationships between satellite and street views, we construct structural and texture controls from satellite images and have designed a cross-view control guided denoising process to enhance the structural and texture fidelity of synthesized panoramic images. Additionally, we extend the traditional satellite-to-street view synthesis task by exploring different data sources, such as text, map data, building height data and multiple-temporal satellite images. In our experiments, we additionally utilize GPT-4o OpenAI (2023) to score synthesized street-view images as a supplement to standard metrics, aiming for a more comprehensive evaluation of the generated results. Experimental results demonstrate that CrossViewDiff excels on three public cross-view datasets, generating realistic and content-consistent images, showcasing outstanding synthesis quality.

The main contributions of this work are summarized as follows:

- We design satellite scene structure estimation and cross-view texture mapping modules to overcome the significant discrepancy between satellite and street views, constructing structure and texture controls for street-view image synthesis.
- We propose a novel cross-view control guided denoising process that incorporates the structure and texture controls via an enhanced cross-view attention module to achieve more realistic street-view panorama synthesis.
- We conduct extensive experiments in street-view image synthesis across a variety of scenes (rural, suburban, and urban), explore additional data sources (e.g. text, maps, multi-temporal images, etc.), and design a GPT-based evaluation metric as a supplement to standard metrics.
- CrossViewDiff outperforms state-of-the-art methods on three public cross-view datasets, achieving an average increase of 9.0% in SSIM, 39.0% in FID, and 35.5% in the GPT-based score.

2 RELATED WORK

2.1 SATELLITE-TO-STREET VIEW SYNTHESIS

Satellite-to-street view synthesis is a challenging task that has been extensively studied. To mitigate the difficulties posed by the large differences across views, many studies explored additional semantic priors to enhance the structure of street-view synthesis results Zhai et al. (2017); Regmi & Borji (2018); Tang et al. (2019); Wu et al. (2022). Zhai et al. Zhai et al. (2017) is a pioneer in this domain that infers the street-view semantic map from the satellite semantic map via a learnable linear transformation. Tang et al. Tang et al. (2019) utilized both the satellite image and the semantic map of street-view image as input to synthesize the target street-view image via image-to-image translation. Although providing a strong structure prior of street-view images, the semantic map is not always available in the actual cross-view synthesis scenarios.

Another group of studies proposed satellite-to-street synthesis methods without using additional semantic information of street-view images, which explored various cross-view projection or transformation methods to provide geometry guidance specifically for panoramic image synthesis Lu et al. (2020); Toker et al. (2021); Shi et al. (2022); Qian et al. (2023). In Lu et al. Lu et al. (2020), a geo-transformation method was proposed for leveraging the height map of satellite view to produce the additional building geometry condition to facilitate street-view panorama synthesis. Toker et al. Toker et al. (2021) applied a polar transformation method proposed by Shi et al. (2019) to cross-view image synthesis and designed a multi-tasks framework in which image synthesis and retrieval are considered jointly. Shi et al. Shi et al. (2022) employed a learnable geographic projection module to learn the geometry relation between the satellite and ground views to facilitate street-view panorama synthesis. Inspired by the success of neural radiance field (NeRF) Mildenhall et al. (2020), Qian et al. Qian et al. (2023) proposed a Sat2Density that can learn a faithful 3D density field as the geometry guidance for panorama synthesis.

In summary, existing studies on satellite-to-street view synthesis are based on generative adversarial networks, with the main aim of improving the structure of synthetic image via semantic or geometric guidance, generating street-view images with low quality and unrealistic textures. By contrast, our study proposes a novel cross-view synthesis method based on Stable Diffusion models Rombach et al. (2022), which designs a cross-view control guided denoising process with a novel cross-view attention module as well as structure and texture controls, generating street-view panoramas with much better perceptual quality and more realistic textures across various scenes.

2.2 DIFFUSION MODELS

In recent computer vision studies, diffusion models Ho et al. (2020) have exhibited remarkable performance in many content creation applications, such as image-to-image translation Saharia et al. (2022a); Li et al. (2023a), text-to-image generation Balaji et al. (2022); Ramesh et al. (2022); Saharia et al. (2022b); Zhang et al. (2024), image enhancement Saharia et al. (2022c); Whang et al. (2022); Gao et al. (2023); Wang et al. (2024), content editing Avrahami et al. (2022); Couairon et al. (2023), and 3D shape generation Luo & Hu (2021); Zeng et al. (2022); Liang et al. (2024); Li et al. (2024b), etc. For traditional denoising diffusion models, images are generated by progressively denoising from random Gaussian noise. For instance, Song et al. Song et al. (2021) proposed denoising diffusion implicit models (DDIM) that reduce the number of denoising steps using an alternative non-Markovian formulation. In latent diffusion models (LDM) Rombach et al. (2022), a variational autoencoder Kingma & Welling (2014) is trained for compressing natural images to a latent space, where the diffusion process will be performed in later stages.

Recently, an increasing number of diffusion models have been proposed for controllable image synthesis Gal et al. (2022); Zhang et al. (2023a); Huang et al. (2023b); Zhao et al. (2023); Ruiz et al. (2023). ControlNet Zhang et al. (2023a) leverages both text and a variety of visual conditions (e.g., sketch, depth map, and human pose) to generate impressive controllable images, which also avoids the need to re-train the entire large model by fine-tuning pre-trained diffusion models and zero-initialized convolution layers. Composer Huang et al. (2023b) integrates global text description with various local controls to train the model from scratch on datasets with billions of samples. Uni-ControlNet Zhao et al. (2023) enables composable control with various conditions using a single model and achieves zero-shot learning on previously unseen tasks. However, these methods utilize similar-view image inputs to control the structure and texture of the synthesis results, resulting in inapplicability to cross-view synthesis tasks.

In addition, several studies have proposed diffusion models for novel view synthesis tasks. For instance, MVDiffusion Tang et al. (2023) proposes a cross-view attention module to generate consistent indoor panoramic images, and Tseng et al. Tseng et al. (2023) utilizes epipolar geometry as a constraint prior to synthesize a consistent video of novel views from a single image. MagicDrive Gao et al. (2024) proposes a street view generation framework that leverages diverse 3D geometry controls (i.e., camera poses, road maps, and 3D bounding boxes) and textual descriptions. However, existing novel view synthesis methods rely on the continuity of image views or camera pose information, which cannot be satisfied in satellite-to-street cross-view settings. Several recent studies have aimed at cross-view synthesis task via diffusion models. Sat2Scene Li et al. (2024c) proposes a novel 3D reconstruction architecture that leverages diffusion models on sparse 3D representations to directly generate 3D urban scenes from satellite imagery. Streetscapes Deng et al. (2024) proposes an autoregressive video diffusion framework and introduces a novel temporal interpolation approach, generating long-range consistent street-view images based on map and height data. However, the task settings of these studies are different from the satellite-to-street view synthesis, and their methods fail to utilize satellite image information to generate realistic street-view textures.

Although diffusion models have achieved promising performance in numerous computer vision applications, few studies have been designed for the challenging satellite-to-street view synthesis task. In this work, we extend the application scenarios of diffusion models to satellite-to-street view synthesis. With both structure and texture controls from the satellite image, our cross-view guided denoising process enables the diffusion model to generate more realistic street-view panoramas.

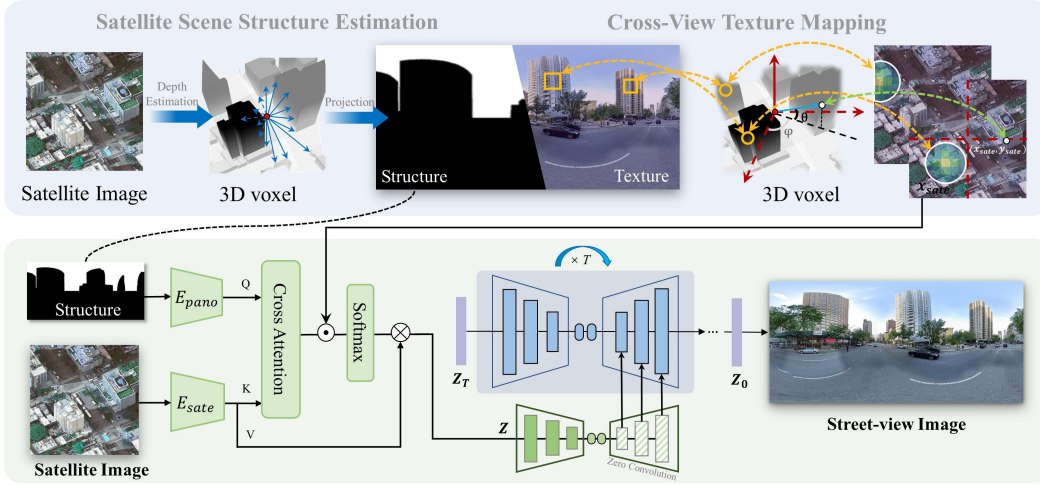


Figure 2: Overview of our proposed CrossViewDiff. First, we create 3D voxels based on a depth estimation method as intermediaries of information across different viewpoints. Subsequently, based on the satellite images and 3D voxels, we establish structural and textural controls for street view synthesis via satellite scene structure estimation and cross-view texture mapping, respectively. Finally, we integrate the above cross-view control information via an enhanced cross-view attention mechanism, guiding the denoising process to synthesize street-view images.

3 METHODS

The goal of satellite-to-street view synthesis is to generate realistic and consistent street-view panoramas from corresponding satellite images. As shown in Figure 2, this paper introduces a novel cross-view synthesis method named CrossViewDiff. In our workflow, we first construct structure and texture controls from satellite images based on the geometric and imaging relationships between satellite and street views. Subsequently, we design a cross-view control guided denoising process via an enhanced cross-view attention module, achieving the synthesis of realistic street-view images.

In the following sections, we first provide a brief introduction to the diffusion model in Section 3.1. In Section 3.2, we discuss the structure and texture controls for cross-view synthesis. In Section 3.3, we describe the cross-view control guided denoising process. In Section 3.4, we detail our strategy for effectively using the GPT model to evaluate the quality of synthesized street-view images.

3.1 PRELIMINARY

Diffusion models are generative models that can generate samples from a Gaussian distribution to match target data distribution by a gradual denoising process Ho et al. (2020). In the forward process, diffusion models gradually add Gaussian noises to a ground truth image x_0 according to a predetermined schedule $\beta_1, \beta_2, \dots, \beta_T$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (1)$$

where x_t is a noised sample with noise level t . The reverse process involves a series of denoising steps, where noise is progressively removed by employing a neural network ϵ_ϕ with parameters ϕ . This neural network predicts the noise ϵ present in a noisy image x_t at step t . The simplified version of the loss function for training the diffusion model is formulated as follows:

$$L_{simple}(\phi, x) = E_{t, \epsilon} \left[\|\epsilon_\phi(x_t, t) - \epsilon\|^2 \right] \quad (2)$$

where t is uniformly sampled from the set $\{1, \dots, T\}$, and x_{t-1} can be reconstructed from x_t by removing the predicted noise:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\phi(x_t, t) \right) + \sqrt{\beta_t} \epsilon \quad (3)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t$ is the cumulative sum of α_t and $\epsilon \sim \mathcal{N}(0, I)$.

3.2 STRUCTURE AND TEXTURE CONTROLS FOR CROSS-VIEW SYNTHESIS

To precisely control the generation of panoramas in cross-view scenarios, it is essential to establish structural and textural information from a street-view perspective based on satellite imagery. Specifically, we start by constructing three-dimensional voxels as intermediaries from the depth estimation results of satellite images. The structural control information is derived from projecting these 3D voxels onto the street-view panorama to obtain scene structure estimates. On the other hand, texture control is achieved through a weight matrix derived from the cross-view mapping relationship based on 3D voxels, representing the response regions on the street view image to different features of the satellite image.

3.2.1 SATELLITE SCENE ESTIMATION FOR STRUCTURE CONTROL

Considering the substantial differences in viewing angles between satellite and street-view modalities, directly extracting contour information from satellite images is challenging. Therefore, we first utilize depth estimation methods to obtain depth results from the satellite perspective Fu et al. (2018); Chen et al. (2019); Yang et al. (2024); Ke et al. (2024). Following this, we convert these depth results into a 3D voxel grid, which serves as an intermediary for scene structure reconstruction. Finally, leveraging the equiangular projection characteristics of street-view panoramas, we establish a mapping from the 3D voxels to the central street view Lu et al. (2020), resulting in a binary map that represents structural information, as shown in Figure 2. This structural information, which includes the positional distribution of significant features (such as buildings, trees, roads, etc.), will further be used as structural control in our diffusion model.

3.2.2 CROSS-VIEW MAPPING FOR TEXTURE CONTROL

Previous methods typically utilize the global texture information of satellite images for panorama synthesis. In contrast, we propose Cross-View Texture Mapping (CVTM), which achieves localized texture control by computing the mapping relationship between each coordinate of the panorama and the satellite image. Based on the 3D voxel grid, we calculate the elevation θ and azimuth ϕ angles from the panoramic image coordinates. For a pixel at $(x_{\text{pano}}, y_{\text{pano}})$ in the panoramic image, the angles are determined as follows:

$$\theta = \frac{\pi}{2} - \frac{y_{\text{pano}} \cdot \pi}{\hat{H}_{\text{pano}}} \quad (4)$$

$$\phi = \frac{x_{\text{pano}} \cdot 2\pi}{\hat{W}_{\text{pano}}} - \pi \quad (5)$$

Here \hat{H}_{pano} and \hat{W}_{pano} denote the height and width of a panoramic image. The calculated angles, θ and ϕ , fall within the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $[-\pi, \pi]$, respectively. According to the two calculated angles, we can determine a ray starting from the center coordinate $(x_{\text{cen}}, y_{\text{cen}})$ of the 3D voxel map. The length of the ray R is the distance from its first intersection with the 3D voxel grid to the center coordinate. Based on the above information, the final mapping coordinates in the satellite image are calculated as follows:

$$x_{\text{sate}} = x_{\text{cen}} + R \cdot \cos(\theta) \cdot \cos(\phi) \quad (6)$$

$$y_{\text{sate}} = y_{\text{cen}} - R \cdot \cos(\theta) \cdot \sin(\phi) \quad (7)$$

Consequently, we establish the pixel-wise mapping relation between each panoramic coordinate $(x_{\text{pano}}, y_{\text{pano}})$ and its corresponding satellite-view coordinate $(x_{\text{sate}}, y_{\text{sate}})$.

In addition, considering the intrinsic errors in cross-view alignment and other factors in complex real-world environment, it is not enough to rely on one-to-one mapping to supplement texture information (the green arrow in Fig 2). The pixels around the mapped points in the satellite images are also valuable texture references that we need to exploit. Consequently, we further design an enhanced satellite texture mapping strategy that leverages the surroundings of the mapped points in the satellite image to enhance the texture details in the street-view image (the orange arrows in Fig 2). This technique utilizes an adaptive re-weighting mechanism based on the distance between the mapped point and other pixels in the satellite image. The values in the weight matrix are calculated as follows:

$$M_j = 1 - \text{sigmoid}(\beta(\|\mathbf{p}^* - \mathbf{p}_j\|_2)) \quad (8)$$

In this formula, \mathbf{p}^* indicates the coordinate $(x_{\text{sat}}, y_{\text{sat}})$ in the satellite image that is mapped from the street-view image according to formula (4)-(7). The \mathbf{p}_j represents each pixel position in the satellite image, where j is an index ranging in $j \in [1, N]$, and N is the number of pixels in the satellite image. The term $\|\cdot\|_2$ is the Euclidean distance. The parameter β controls the rate of change in the sigmoid function. The weight value M_j indicates the importance of \mathbf{p}_j to the mapped point \mathbf{p}^* , which will be higher if \mathbf{p}_j is close to \mathbf{p}^* , thus enhancing the overall realism and coherence of the street-view images. Consequently, we have obtained the weight matrix M , which reflects the texture mapping relationship between satellite and street-view images.

3.3 CROSS-VIEW CONTROL GUIDED DENOISING PROCESS

Based on satellite scene estimation, we obtain binary maps to serve as structural controls for the street-view images. Utilizing cross-view mapping, we derive weight matrices to act as texture controls for the street-view images. Based on the characteristics of the structural and textural control information, we design an enhanced cross-view attention module to integrate both types of information, guiding the subsequent denoising process.

In our enhanced cross-view attention module, let $Q \in \mathbb{R}^{h_p \times w_p}$ denote the Query feature from the panoramic binary map S_{pano} , $K \in \mathbb{R}^{h_s \times w_s}$ denote the Key feature from the input satellite image I_{sat} , and $V \in \mathbb{R}^{h_s \times w_s}$ denote the Value feature, which contain detailed texture information from the satellite image. Here, $h_p \times w_p$ and $h_s \times w_s$ represent the resolution of the panorama and satellite feature map, respectively. Moreover, E_{sat} and E_{pano} denote the satellite and panoramic encoders. W_q , W_k and W_v are projection matrices. The definitions of Q , K , V are as follows:

$$Q = W_q(E_{\text{pano}}(S_{\text{pano}})), \quad K = W_k(E_{\text{sat}}(I_{\text{sat}})), \quad V = W_v(E_{\text{sat}}(I_{\text{sat}})), \quad (9)$$

The process begins with the computation of an affinity matrix $A \in \mathbb{R}^{h_p w_p \times h_s w_s}$, reflecting the interaction between Q and K . Following this, the weight matrix derived from the previous module is down-sampled to $M \in \mathbb{R}^{h_p w_p \times h_s w_s}$ and applied to each pixel within the satellite image to emphasize relevant features. This selective enhancement is crucial for the subsequent fusion of the detailed texture information from the satellite image into the panoramic feature $F_{\text{pano}} \in \mathbb{R}^{h_p \times w_p}$. The enhanced cross-view attention mechanism is formulated as follows:

$$\mathbf{z} = \text{softmax}(A \odot M) \cdot V \quad (10)$$

In these expressions, \odot denotes element-wise multiplication, where the weight matrix M is applied to the affinity matrix (A) to obtain the reweighted affinity matrix (A'), emphasizing the connection between the most relevant pixels.

The output \mathbf{z} , generated at each step, is ingeniously reincorporated into the network as a pivotal conditional element. By employing \mathbf{z} as a dynamic conditional catalyst within our cross-view diffusion architecture, we ensure that each step of the denoising process is informed by the evolving latent representation, thereby enabling a controlled and gradual transition from z_t to z_0 . This process is meticulously orchestrated by a cross-view control guided denoising process, which integrates structural and textural knowledge extracted from I_{sat} into the refinement of the final latent feature \mathbf{z} , subsequently decoded through Stable Diffusion’s latent space decoder \mathcal{D} to achieve the generated street-view panorama I_{pano} .

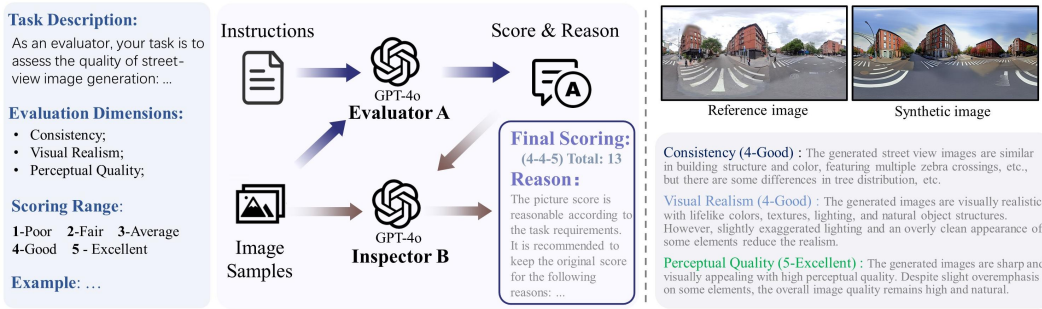


Figure 3: The overall process for automated evaluation using GPT-4o. Instructions are meta-prompts that include a task description, scoring criteria, scoring range, and scoring examples. Then we use a GPT-4o as Evaluator A to provide initial scores and reasons based on the input prompts and image samples. Finally, the scores are combined with the image samples for a secondary evaluation by another GPT-4o as Inspector B, who assesses the score’s appropriateness and determines the final score.

3.4 GPT-BASED EVALUATION METHOD FOR CROSS-VIEW SYNTHESIS

Cross-modal satellite-to-ground synthesis requires measuring both the consistency and realism of generated images. Traditional metrics like SSIM Wang et al. (2004) and FID Heusel et al. (2017) generally focus on single dimensions of similarity or realism, providing incomplete evaluations. Inspired by the use of large multimodal models for synthetic image scoring Cho et al. (2023); Huang et al. (2023a); Wu et al. (2024); Zhang et al. (2023b), we design a new evaluation process based on GPT-4o, as shown in Figure 3. This approach enables comprehensive and interpretable assessments of synthesized street-view images, aligning more closely with human judgment standards.

Firstly, we design three key evaluation dimensions for cross-view synthesized images: Consistency, Visual and Structural Realism, and Perceptual Quality. We adopted a rating scheme, establishing a 5-level rating system with scores ranging from 1 (poor) to 5 (excellent).

Consistency: This dimension evaluates the alignment of the content in synthesized images with real street-view images, including the structure and texture of buildings, the layout of roads, and the similarity of other significant landmarks, measuring the content consistency of the synthesized street-view images.

Visual Realism: This evaluates the visual effect and structural reasonableness of the generated images, including the realism of color, shape, and texture, as well as the the structural integrity, assessing whether they look like real street-view images.

Perceptual Quality: This evaluates the overall perceptual quality of the generated images, including aspects such as image clarity, noise level, and visual comfort, measuring the quality of the generated images.

To achieve more effective GPT scoring, we employed Chain-of-Thought (CoT) and In-Context Learning (ICL) strategies Alayrac et al. (2022); Zhang et al. (2023c); Brown et al. (2020); Peng et al. (2024) to enhance its stability and effectiveness. Firstly, we provided GPT-4o with a small number of effective human-scored examples from multiple users, enabling the model to effectively learn human scoring patterns. Secondly, by enabling the large model to explain the reasoning behind its scores, we have introduced an element of internal reflection to the evaluation process. Additionally, we used GPT-4o to act as Evaluator A and Inspector B. After receiving the initial scores and reasons from Evaluator A, Inspector B will assess the reasonableness of these scores and make the final scoring decision. If the scores are deemed reasonable, they will be retained; otherwise, Inspector B will provide new scoring results and justifications.

To validate the effectiveness of GPT-based scoring, we invited ten human users to perform the same scoring task and measured the consistency between their scores and those generated by GPT. We provided thorough training to the users to ensure they fully understood the satellite-to-street view generation task. The users’ scoring tasks and schemes were aligned with the GPT scoring. We

ensured that each generated image was scored by at least two human users. Due to the large volume of cross-view datasets and the cost of both user and GPT scoring, we randomly sampled 1000 images from the evaluation sets of each dataset for assessment. In addition to our method, we selected the best comparative results from GAN and diffusion methods for evaluation. A total of 9000 images were used for user scoring, and we measured the agreement between these scores and the GPT scores.

4 EXPERIMENTS

In this section, we first introduce the three datasets used in this study and the experimental setting. Next, we conduct both qualitative and quantitative comparisons of CrossViewDiff with state-of-the-art cross-view synthesis methods. Following this, we perform ablation studies to evaluate the effectiveness of each module. Additionally, we explore street-view synthesis tasks using additional data sources. Finally, we discuss the limitations of our method.

4.1 DATASET

In our experiments, we used three popular cross-view datasets to evaluate the synthesis results, i.e., CVUSA Zhai et al. (2017), CVACT Liu & Li (2019) and OmniCity Li et al. (2023c). These three datasets encompass rural, suburban, and urban scenes, providing a robust benchmark for comprehensively evaluating the performance of satellite-to-street view synthesis. Furthermore, in addition to the original satellite imagery and building height data provided by OmniCity, we supplemented multimodal data including text, maps, and multi-temporal satellite imagery, providing data support for street-view synthesis tasks using additional multimodal data sources.

CVUSA Zhai et al. (2017) is a standard large-scale cross-view benchmark, primarily featuring rural scenes such as roads, grasslands, and forests. This dataset comprises centrally aligned satellite and street-view images collected from various locations across the United States, which is randomly split into training and test sets in an 8:2 ratio.

CVACT Liu & Li (2019) is a widely used cross-view dataset that includes satellite and street-view images from Canberra, Australia. This dataset mainly consists of suburban scenes with relatively low buildings and open views. Unlike CVUSA dataset, the training and test sets of CVACT dataset are divided by region.

OmniCity Li et al. (2023c) is an urban cross-view dataset that includes street-view and satellite images from New York, USA. The primary scenes in OmniCity consist of dense urban buildings, and street-view images that are heavily obstructed by trees or vehicles will be filtered out. OmniCity is divided into training and test data by region.

Additionally, the orientation towards the north in both street view and satellite imagery is a critical attribute for cross-view datasets. In all three datasets, the north direction in satellite images is at the top of the image. In CVUSA Zhai et al. (2017) and CVACT Liu & Li (2019), the north direction in street-view images is in the center column, while in OmniCity Li et al. (2023c), it is in the first column.

4.2 EXPERIMENTAL SETTING

We implement CrossViewDiff based on the ControlNet Zhang et al. (2023a) framework, incorporating the pre-trained Stable Diffusion Rombach et al. (2022) v1.5 model. The diffusion decoder is configured in an unlocked state and the classifier-free guidance Ho & Salimans (2022) scale is established at 9.0. For the final inference sampling, we adopt $T = 50$ as the sampling step, consistent with the DDIM Song et al. (2021) strategy. The entire training process is performed on eight NVIDIA A100 GPUs, with a batch size of 128, spanning a total of 100 epochs. Our depth estimation method employs Marigold Ke et al. (2024) and is fine-tuned on the OmniCity dataset, which provides elevation information. We conduct our experiments at a resolution of 1024×256 on the CVUSA Zhai et al. (2017) and 1024×512 on OmniCity Li et al. (2023c) and CVACT Liu & Li (2019).

We compared our method with several state-of-the-art cross-view synthesis methods on the three datasets, including GAN-based methods such as Sate2Ground Lu et al. (2020), CDTE Toker et al. (2021), S2SP Shi et al. (2022), and Sat2Density Qian et al. (2023), as well as diffusion models for image transformation control like ControlNet Zhang et al. (2023a) and Instruct pix2pix (Instr-p2p) Brooks et al. (2023). For Sat2Density Qian et al. (2023), we follow their original setup, i.e., the lighting hints are determined based on the average values of the sky histograms obtained from random selections. For diffusion-based methods (ControlNet and instr-p2p), we use a pre-trained model consistent with that of CrossViewDiff and maintain the same sample steps. Note that all comparison methods are conducted according to their optimal experimental settings.

Following previous studies Lu et al. (2020); Toker et al. (2021); Regmi & Borji (2018), we used common metrics such as SSIM Wang et al. (2004), SD, and PSNR to evaluate the content consistency of synthesized images, and FID Heusel et al. (2017) and KID Bińkowski et al. (2018) to assess image realism. Furthermore, in Section 4.3.2, we use GPT-4o to evaluate the synthesized street view images across three dimensions: consistency, visual realism, and perceptual quality.

4.3 COMPARISON WITH STATE-OF-THE-ART METHODS

4.3.1 QUANTITATIVE AND QUALITATIVE EVALUATION

Table 1: Quantitative comparison of different methods on CVUSA Li et al. (2023c) and CVACT Liu & Li (2019) datasets in terms of various evaluation metrics.

Method	CVUSA					CVACT				
	SSIM (↑)	SD (↑)	PSNR (↑)	FID (↓)	KID (↓)	SSIM (↑)	SD (↑)	PSNR (↑)	FID (↓)	KID (↓)
Sate2Ground Lu et al. (2020)	0.294	15.48	12.634	52.42	0.036	0.392	15.09	13.038	55.61	0.079
CDTE Toker et al. (2021)	0.283	15.24	13.815	28.35	0.028	0.370	15.52	13.707	57.00	0.064
S2SP Shi et al. (2022)	0.319	15.73	13.689	27.31	0.021	0.368	15.86	13.974	65.38	0.064
Sat2Density Qian et al. (2023)	0.339	15.73	14.229	41.43	0.036	0.387	16.09	14.271	47.09	0.038
ControlNet Zhang et al. (2023a)	0.277	15.22	11.182	44.63	0.044	0.340	15.36	12.150	47.15	0.019
Instruct pix2pix Brooks et al. (2023)	0.255	15.76	10.664	68.75	0.077	0.392	15.64	13.123	57.74	0.049
Ours	0.371	16.31	12.000	23.67	0.018	0.412	16.29	12.411	41.94	0.041

Table 2: Quantitative comparison of different methods on OmniCity Li et al. (2023c) dataset in terms of various evaluation metrics.

Method	OmniCity				
	SSIM (↑)	SD (↑)	PSNR (↑)	FID (↓)	KID (↓)
Sate2Ground Lu et al. (2020)	0.290	14.38	12.430	75.22	0.053
CDTE Toker et al. (2021)	0.294	14.47	11.594	122.29	0.141
S2SP Shi et al. (2022)	0.294	14.61	12.748	84.00	0.088
Sat2Density Qian et al. (2023)	0.316	14.73	13.661	87.90	0.072
ControlNet Zhang et al. (2023a)	0.297	14.64	10.703	59.99	0.056
Instruct pix2pix Brooks et al. (2023)	0.291	14.03	10.363	64.89	0.087
Ours	0.353	15.17	11.127	42.01	0.033

We provide the quantitative results on the rural CVUSA and suburban CVACT datasets in Table 1. Compared to the state-of-the-art method for cross-view synthesis (Sat2Density), our method achieved significant improvements in SSIM Wang et al. (2004) and FID Heusel et al. (2017) scores by 9.44% and 42.87% on CVUSA, respectively. Similarly, enhancements of 6.46% and 10.94% in SSIM and FID were observed on CVACT. Visual results from Figure 4 suggest that GAN-based cross-view methods tend to produce excessive artifacts and blurriness. While diffusion-based approaches like ControlNet Zhang et al. (2023a) and Instr-p2p Brooks et al. (2023) can generate highly realistic street views, they often lack content relevancy with the Ground Truth. In contrast, our method benefits from structure and texture controls, effectively capturing satellite-view information to generate realistic images that are more consistent with the Ground Truth street-view images, including buildings, trees, green spaces, and roads.

In the urban OmniCity dataset, our CrossViewDiff also demonstrates excellent performance compared to the most advanced methods, as shown in Table 2. Compared with the state-of-the-art (Sat2Density Qian et al. (2023)), our approach achieves significant improvements in SSIM Wang et al. (2004) and FID Heusel et al. (2017) by 11.71% and 52.22%, respectively. The visual results



Figure 4: Qualitative comparison of synthesis results on CVUSA Zhai et al. (2017), CVACT Liu & Li (2019) and OmniCity Li et al. (2023c), respectively. The comparison includes the synthesis results of Sat2Density Qian et al. (2023), ControlNet Zhang et al. (2023a), Instr-p2p Brooks et al. (2023), and our method. The results indicate that our method generates street views that are more realistic, consistent, and of higher quality compared to other methods.

from the last three rows of Figure 4 demonstrate that our method effectively maintains good performance in synthesized street view images of urban scenes, such as more realistic and consistent building contours and colors. Extensive experimental results demonstrate that our CrossViewDiff outperforms existing methods and achieves excellent results for street-view image synthesis across various scenes, including rural, suburban and urban environments.

4.3.2 GPT-BASED EVALUATION

Beyond conventional similarity and realism metrics, we also leverage the powerful visual-linguistic capabilities of existing MLLM large models to design a GPT-based scoring method for evaluating synthetic images. As shown in Figure 5, GPT can provide scores across multiple dimensions along with the corresponding reasons for the scores. The description of the scoring reasons by GPT enhances the interpretability of the metric scores. As described in section 3.4, a subset of the dataset (9K pairs of images) was evaluated by both human users and GPT. By calculating the similarity between each user rating and the GPT score, the results, as shown in Table 3, demonstrate that GPT-based scoring performs well in aligning with human ratings across multiple metrics, with an average similarity exceeding 80%. This highlights the fact that GPT-based scoring is very close to human preferences and can effectively evaluate synthetic street-view images.

Moreover, as illustrated in Table 4 and Figure 6, our method significantly outperforms other GAN-based and diffusion-based generation methods in the three evaluation dimensions of Consistency,

Visual Realism, and Perceptual Quality. This also indicates that the street-view images synthesized by our method are more aligned with the requirement of human users, which aids in subsequent applications such as immersive scenes and virtual reality tasks.





Street-view Image	Score & Reason
 <p>Sat2Density</p>	<p>Sat2Density</p> <p>Consistency (2) : The generated image does not closely match the reference ... structure and layout of the road have certain similarities ... the overall texture and details differ significantly ...</p> <p>Visual Realism (1) : The generated image lacks the realistic details ... textures and colors are less defined ... difficult to distinguish between the road and the buildings...</p> <p>Perceptual Quality (1) : The overall quality of the generated image is low, with blurriness and a lack of sharpness ... The image appears distorted and lacks the clarity ...</p>
 <p>Instruct pix2pix</p>	<p>Instruct pix2pix</p> <p>Consistency (1) : The generated image does not match the reference image in most recognizable features, structures, or layout ... It appears to be a completely different street view ...</p> <p>Visual Realism (4) : The generated image shows a high degree of visual realism, with lifelike textures and details in sky, roads, and vehicles ...</p> <p>Perceptual Quality (3) : The perceptual quality of the image is average ... is relatively clear, but the lack of fine detail of buildings ... detract from the overall visual appeal ...</p>
 <p>Ours</p>	<p>Ours</p> <p>Consistency (4) : The generated image maintains a similar urban environment with street intersections, buildings ... slight differences in the specific placement and architecture of the buildings ...</p> <p>Visual Realism (4) : ... provides a good level of visual realism ... The lighting and textures appear natural and lifelike ... However, the branches of the trees have a certain sense of unreality ...</p> <p>Perceptual Quality (4) : It is clear and visually appealing, with well-defined elements such as buildings and trees ... minor issues with sky and some slight blurring ...</p>
 <p>GT</p>	

Figure 5: An example of GPT-based evaluation. Given a synthesized street-view image and the corresponding Ground Truth, GPT-based evaluation can provide scores across multiple dimensions and the corresponding reasons for the scores.

Table 3: Average similarity between human user ratings and GPT ratings.

Evaluation Metrics	Average Similarity
Consistency	0.810
Visual Realism	0.816
Perceptual Quality	0.743
Total Score	0.801

Figure 6: GPT-based evaluation results.

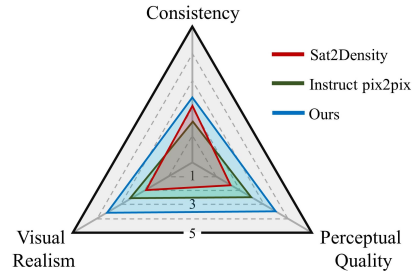


Table 4: Evaluation results of street view synthesis based on GPT-4o. The scores range from 1 (poor) to 5 (excellent), presenting the average score across three datasets. Our method significantly outperforms other methods in terms of the three evaluation dimensions and the total score.

Method	Consistency	Visual Realism	Perceptual Quality	Total Score
Sat2Density Qian et al. (2023)	2.07	2.05	1.74	7.91
Instruct pix2pix Brooks et al. (2023)	1.67	2.75	2.61	9.79
Ours	2.32	3.66	3.66	13.27

4.3.3 PANORAMA CONTINUITY EVALUATION

For street-view panorama synthesis, another important evaluation factor is the continuity between the left and right sides of the image. As illustrated in the qualitative results in Figure 7, both GAN-based and diffusion-based methods produce synthesis results with apparent boundary lines, as they treat panorama synthesis as a general image synthesis task. In contrast, our method constructs structural controls from a continuous scene composed of 3D voxels projected onto panoramic street

views, allowing seamless integration at the left and right boundaries. For texture controls, the texture mapping features at the left and right positions of the street views are derived from proximate and continuous positions on the satellite image. Owing to these continuous structural and textural constraints, our method produces panoramic images with excellent 360° coherence.

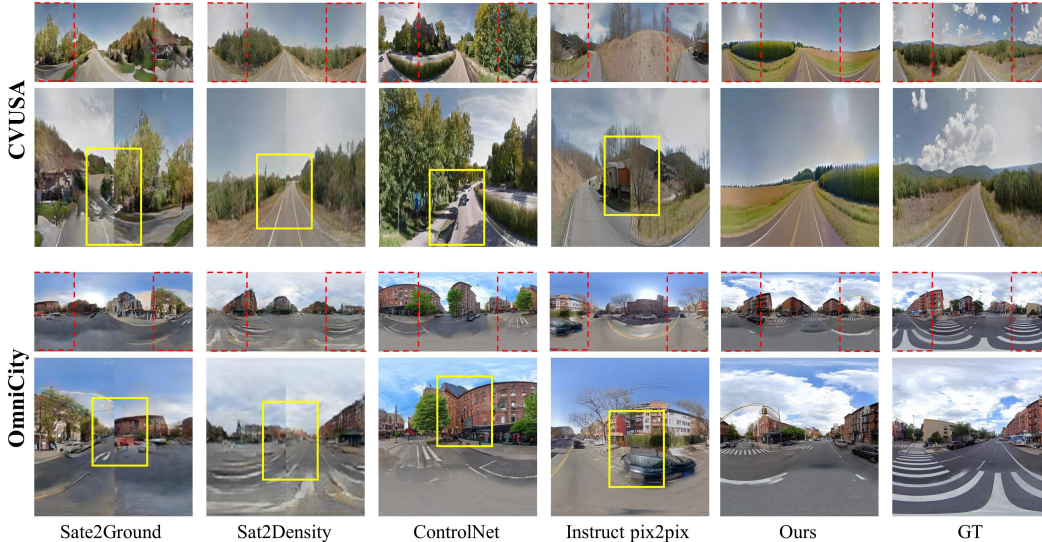


Figure 7: Qualitative results of the panorama continuity evaluation on CVUSA Zhai et al. (2017) and OmniCity Li et al. (2023c), respectively. By stitching the right 90° of the synthesis panorama to the left side of the image, our method demonstrates excellent consistency in texture and structure compared to other methods.

4.4 ABLATION STUDY

In our ablation study, we first assessed the effectiveness of our structure and texture control modules. As shown in the first two rows of each dataset of Table 5, using structural information derived from satellites as input proved effective, achieving improvements across multiple metrics such as SSIM Wang et al. (2004), FID Heusel et al. (2017), and KID Biríkowski et al. (2018). The last two rows of each dataset show the results of using direct Cross-Attention to incorporate global textures (w/o CVTM) and our Cross-View Texture Mapping (w/ CVTM) methods. Compared to the direct incorporation global textures, the approach guided by cross-view mapping relationships effectively assigns local textures from corresponding satellite regions to the appropriate locations in street-view images. Figure 8 presents qualitative ablation results on CVUSA Zhai et al. (2017) and OmniCity Li et al. (2023c), where structural control contributes to consistent content distribution, and texture control enhances the consistency of generated textures in buildings and forests.

Table 5: Quantitative ablation for different types of controls on CVUSA Zhai et al. (2017) and OmniCity Li et al. (2023c), including Structure, Texture (w/o CVTM), and Texture (w/ CVTM).

Datasets	Structure	Texture (w/o CVTM)	Texture (w/ CVTM)	SSIM (↑)	SD (↑)	PSNR (↑)	FID (↓)	KID (↓)
CVUSA	✓			0.277	15.22	11.182	44.63	0.044
	✓			0.312	15.30	10.358	41.19	0.039
	✓	✓		0.283	15.65	10.913	33.51	0.020
	✓		✓	0.371	16.31	12.000	23.67	0.018
OmniCity	✓			0.297	14.64	10.703	59.99	0.056
	✓			0.309	14.54	11.417	43.06	0.042
	✓	✓		0.345	14.73	10.899	64.33	0.059
	✓		✓	0.353	15.17	11.127	42.01	0.033

Additionally, as the intermediaries for constructing both structural and textural controls, the 3D voxels derived from satellite depth estimation results significantly impact the accuracy of cross-view

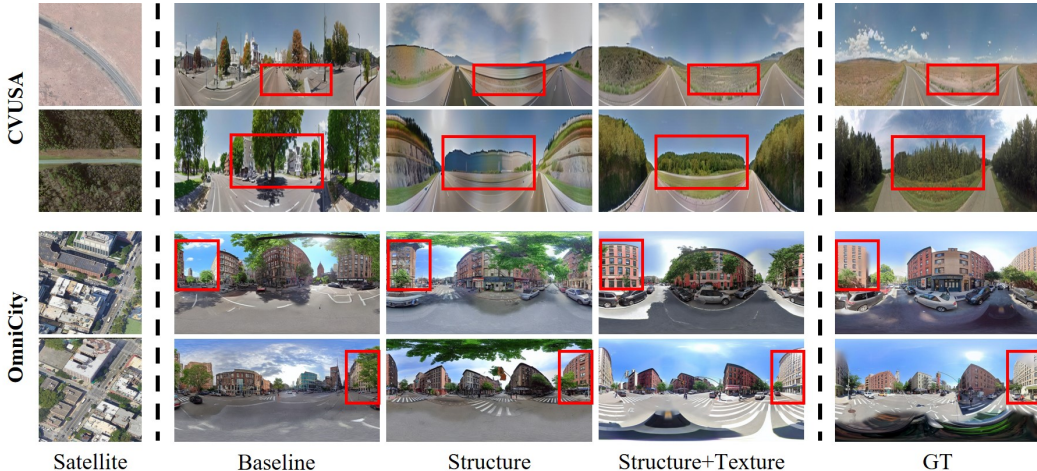


Figure 8: Qualitative ablation results on CVUSA Zhai et al. (2017) and OmniCity Li et al. (2023c). In the synthesis results, the first column represents the baseline without any structure or texture controls, the second column represents using only structure constraints, and the third column represents using both structure and texture (w/ CVTM) controls.

Table 6: Ablation results for varying depth estimations on CVUSA Zhai et al. (2017) and OmniCity Li et al. (2023c) datasets. The impact of adjusted depth results on experimental metrics is minimal.

Method	OmniCity					CVUSA				
	SSIM (\uparrow)	SD (\uparrow)	PSNR (\uparrow)	FID (\downarrow)	KID (\downarrow)	SSIM (\uparrow)	SD (\uparrow)	PSNR (\uparrow)	FID (\downarrow)	KID (\downarrow)
Ours ($\times 0.9$)	0.350	15.10	11.111	43.58	0.034	0.365	16.30	11.943	24.11	0.019
Ours ($\times 1.1$)	0.349	15.11	11.104	44.76	0.037	0.368	16.29	11.950	23.13	0.019
Ours	0.353	15.17	11.127	42.01	0.033	0.371	16.31	12.000	23.67	0.018
Δ	0.004	0.07	0.023	2.75	0.004	0.006	0.02	0.057	0.98	0.001

controls. Therefore, the precision of satellite depth estimation directly influences the effectiveness of these controls. To simulate depth estimation inaccuracies, we apply scaling factors (0.9 and 1.1) to the depth estimation results before generating street-view images, as detailed in Table 6. The experimental results indicate that while our method relies on depth estimation, the stability of the model’s output remains high, with minimal fluctuation in performance metrics.

4.5 EXPERIMENTAL RESULTS USING ADDITIONAL DATA SOURCES

In this section, we provide more experimental results of real-world application scenarios using additional data sources. In addition to the satellite images, other inputs such as textual data, building height data, and public map data (e.g. OpenStreetMap¹) can also be used for generating street-view images. In this study, we explored the synthesis of street-view images using multiple data sources on the OmniCity Li et al. (2023c) dataset and analyzed their impacts. Based on OmniCity street-view images, we generate corresponding text prompts of street-views images using the CLIP Radford et al. (2021) model, and supplement the corresponding historical satellite imagery and OSM map data based on the street view capture locations.

As shown in Figure 9, textual data can provide some global information about the scene, but its lack of detail and specificity results in visually unrealistic images. OSM (OpenStreetMap) data offers semantic features of different areas, such as roads, buildings, and parks. These semantic features aid in generating street-view images with consistent semantic content. However, when using only OSM data, the structure and texture of the synthesized street view images still show a certain gap compared to real images. Building height data provides the outlines of buildings, and street-view images synthesized using this data show consistent building contours but lack texture

¹<https://www.openstreetmap.org/>

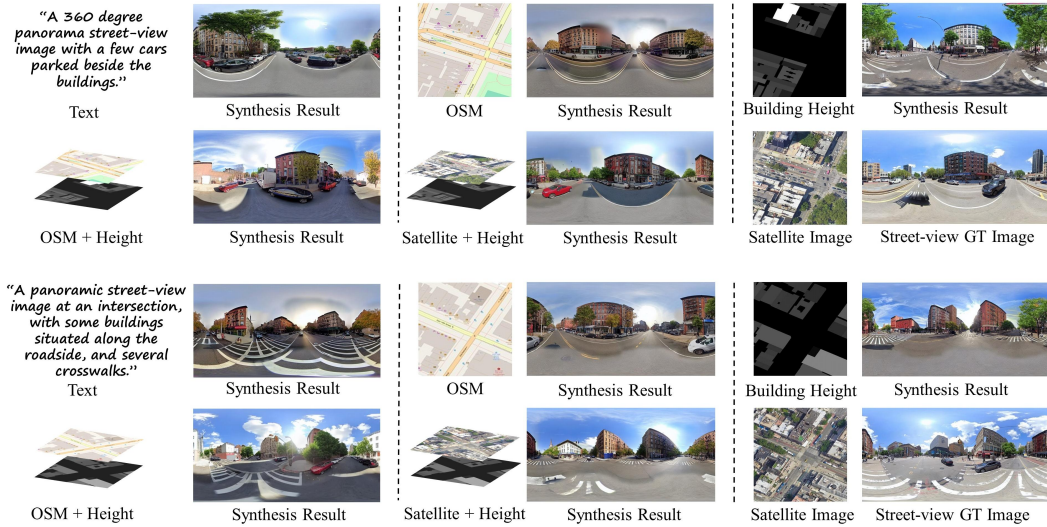


Figure 9: Qualitative comparison of different input types on the OmniCity Li et al. (2023c) dataset. Using satellite image and building height as input achieves the best results in all cases.

and detail. Combining OSM and building height data for street view synthesis perform well in terms of semantics and structure. However, there are still deficiencies in texture details, such as building colors. Combining satellite imagery and building height data yields street-view images that are optimal in both structure and texture, visually closest to real street views. Table 7 provide the quantitative results obtained from different types of input data. Due to the rich texture information in satellite images, our CrossViewDiff achieved SSIM Wang et al. (2004) and FID Heusel et al. (2017) scores of 0.361 and 37.89, respectively, representing improvements of 4.6% and 17.6% compared to the results synthesized using OSM and building height data as inputs.

Table 7: Quantitative comparison of different types of input data on the OmniCity dataset. Using satellite image and building height as input data achieves optimal performance, with a significant improvements compared with other input cases.

Input data	SSIM (\uparrow)	SD (\uparrow)	PSNR (\uparrow)	FID (\downarrow)	KID (\downarrow)
Text	0.298	14.54	11.131	82.37	0.069
OSM	0.294	14.67	10.741	43.26	0.034
Building height	0.327	14.65	11.422	47.94	0.044
OSM + Building height	0.345	14.79	11.748	45.98	0.039
Satellite + Building height	0.361	15.21	11.512	37.89	0.027

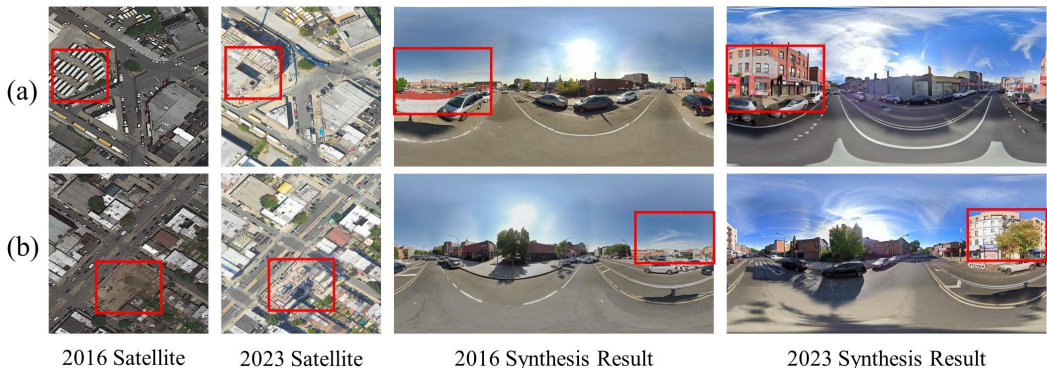


Figure 10: Visualization results of street-view synthesis from satellite images taken at different times. The areas highlighted in red indicate regions where terrain changes have occurred over time.

Next, we explored the results of synthesizing street-view images using satellite imagery data from different years. As shown in Figure 10, significant changes in terrain features over time can also be observed in our synthesized street-view images, such as the transformation of parking lots or vacant lots into buildings within the areas highlighted in red. Given the relatively recent widespread adoption of street-view imaging compared to the earlier availability of remote sensing satellite imagery, our effective satellite-to-street-view synthesis method unveils historical scenes from earlier times, offering practical application value.

4.6 LIMITATION ANALYSIS

Despite the above advantages, street-view images generated by CrossViewDiff still have several limitations. Although we fused features rich in structural and textural information based on satellite image, the gap between the two viewpoints is still large, and Stable Diffusion is more capable of creating additional details that do not actually exist. Figure 11 provides some typical failure cases obtained by CrossViewDiff. For satellite and street-view images that were not taken at the same season, even though the synthetic street-view image is consistent with the satellite’s features, it may not be consistent with the ground truth. Besides, in less constrained regions of the image such as the sky, the synthesis result is somewhat different from GT and has a certain amount of color shifting, resulting in the relatively low PSNR to some extent. Moreover, due to the presence of moving objects such as pedestrians and vehicles in the scene, achieving consistency in cross-view synthesis results remains challenging.

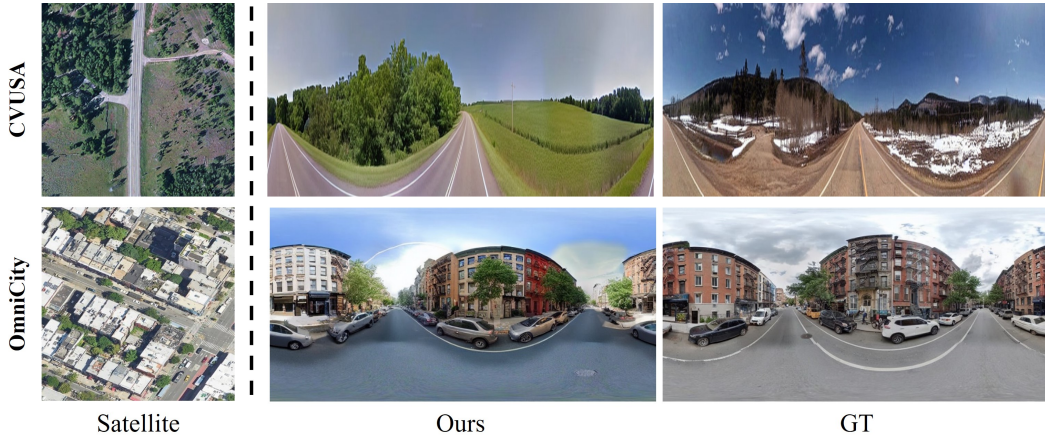


Figure 11: Typical failure cases of our method. The first row of images shows that as the satellite and street-view images provided in the dataset were not taken at the same season, the synthetic image may not be consistent with the ground truth even if it is consistent with the satellite’s features. The second row shows a significant discrepancy in the sky areas of the synthesized street views, as sky region information cannot be obtained from satellite images. Additionally, vehicles and other moving objects pose significant challenges to cross-view synthesis.

5 CONCLUSION

In this work, we have proposed CrossViewDiff, a cross-view diffusion model to synthesize a street-view panorama from a given satellite image. The core of our diffusion model is a cross-view control guided denoising process that incorporates the structure and texture controls constructed by satellite scene structure estimation and cross-view texture mapping via an enhanced cross-view attention module. Qualitative and quantitative results show that our method generates street-view panoramas with better consistency and perceptual quality as well as more realistic structures and textures compared with the state-of-the-art. We believe that this paper motivates new ideas and inspirations for large-scale city simulation and 3D scene reconstruction. In our future work, we will further explore the fusion of more types of multimodal data including textual data, map data, 3D data, and multi-temporal satellite imagery to enhance the quality and realism of the synthesized street-view images.

We also plan to extend our method to more cities and improve our methods for more complex application scenes such as urban planning, virtual tourism, and intelligent navigation.

DECLARATIONS

Data Availability The datasets used this study can be accessed from: 1) CVUSA: <https://mvr1.cse.wustl.edu/datasets/cvusa>. 2) CVACT: <https://github.com/Liumouliu/OriCNN>. 3) OmniCity: <https://city-super.github.io/omnicity>.

Code Availability The implementation code and models related to the paper will be released at <https://opendatalab.github.io/CrossViewDiff>.

REFERENCES

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18208–18218, June 2022.
- Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.
- 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*, pp. 18392–18402, 2023.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. 2019.
- Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.
- Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023.
- Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.
- Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2002–2011, 2018.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

- Ruiyuan Gao, Kai Chen, Enze Xie, HONG Lanqing, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, 2024.
- Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10021–10030, 2023.
- 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. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.
- Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023a.
- Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *ICML*, 2023b.
- Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.
- Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023.
- Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961, June 2023a.
- Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16719–16729, 2024a.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.
- Ming Li, Pan Zhou, Jia-Wei Liu, Jussi Keppo, Min Lin, Shuicheng Yan, and Xiangyu Xu. Instant3d: Instant text-to-3d generation. *International Journal of Computer Vision*, pp. 1–17, 2024b.
- Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omniscity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17397–17407, June 2023c.
- Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R Oswald. Sat2scene: 3d urban scene generation from satellite images with diffusion. *arXiv preprint arXiv:2401.10786*, 2024c.

- Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pp. 1–21, 2024.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5624–5633, 2019.
- 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*, pp. 859–867, 2020.
- Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.
- 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.
- R OpenAI. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023.
- Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.
- 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 (ICCV)*, pp. 3683–3692, October 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3501–3510, 2018.
- 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*, pp. 10684–10695, 2022.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.
- Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450393379.
- 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, 2022b.

- Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022c.
- Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.
- Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10009–10022, 2022.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2417–2426, 2019.
- Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 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*, pp. 6488–6497, 2021.
- Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhub Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16773–16783, 2023.
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pp. 1–21, 2024.
- 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.
- Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16293–16303, 2022.
- Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 2022.
- Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22227–22238, 2024.
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.
- Junyan Ye, Jun He, Weijia Li, Zhutao Lv, Jinhua Yu, Haote Yang, and Conghui He. Skydiffusion: Street-to-satellite image synthesis with diffusion models and bev paradigm, 2024a. URL <https://arxiv.org/abs/2408.01812>.
- Junyan Ye, Qiyang Luo, Jinhua Yu, Huaping Zhong, Zhimeng Zheng, Conghui He, and Weijia Li. Sg-bev: Satellite-guided bev fusion for cross-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27748–27757, 2024b.

- Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network, 2024c. URL <https://arxiv.org/abs/2408.05475>.
- Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- Kaiduo Zhang, Muyi Sun, Jianxin Sun, Kunbo Zhang, Zhenan Sun, and Tieniu Tan. Open-vocabulary text-driven human image generation. *International Journal of Computer Vision*, pp. 1–19, 2024.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023a.
- Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023b.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023c.
- Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.