Geometry-Guided Cross-View Diffusion for One-to-Many Cross-View Image Synthesis

Tao Jun Lin[1], Wenqing Wang[2], Yujiao Shi[3],
Akhil Perincherry[4], Ankit Vora[4] and Hongdong Li[1]
[1]The Australian National University [2]University of Surrey
[3]ShanghaiTech University [4]Ford Motor Company
taojun.lin@anu.edu.au, shiyj2@shanghaitech.edu.cn

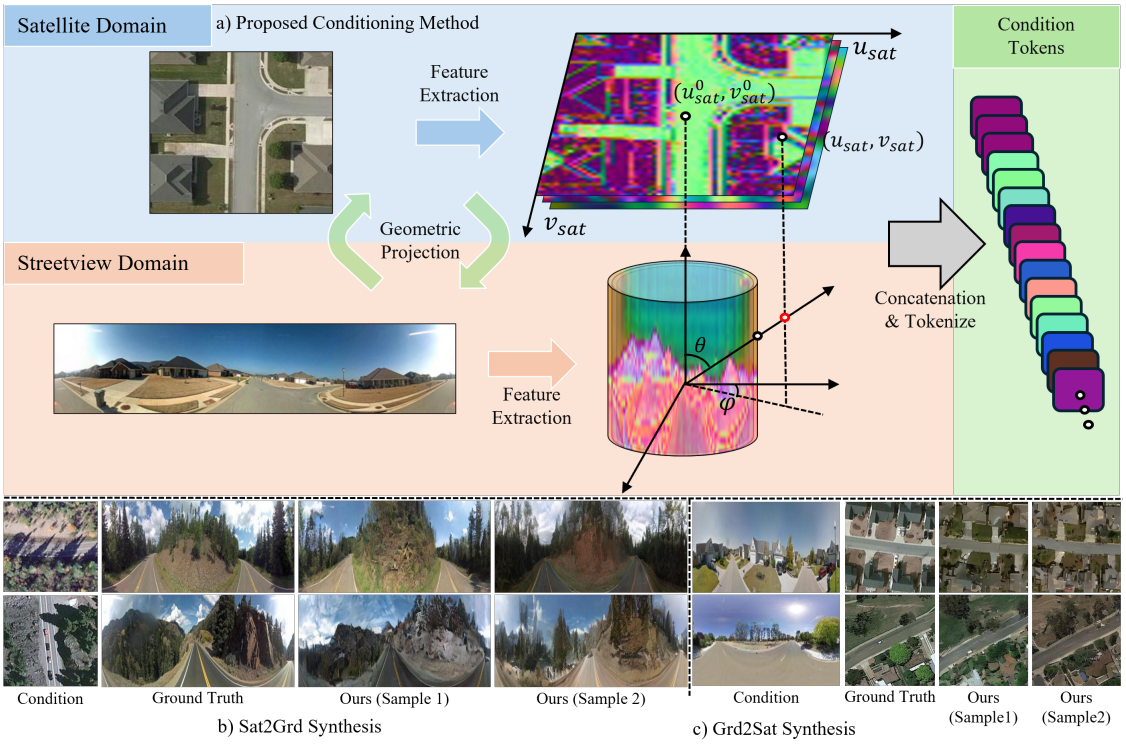


Figure 1. Our proposed Geometry-Guided Conditioning method for cross-view image synthesis (a) and visualization examples generated by our proposed Geometry-guided Cross-view Diffusion. On the bottom left (b) are images generated from our **Sat2Grd** model, on the bottom right (c) are image generated from our **Grd2Sat** model.

Abstract

This paper presents a novel approach for cross-view synthesis aimed at generating plausible ground-level images from corresponding satellite imagery or vice versa. We refer to these tasks as satellite-to-ground (Sat2Grd) and ground-to-satellite (Grd2Sat) synthesis, respectively. Unlike previous works that typically focus on one-to-one generation, producing a single output image from a single input image, our approach acknowledges the inherent one-to-many nature of the problem. This recognition stems from the challenges posed by differences in illumination, weather conditions, and occlusions between the two views. To effectively model this uncertainty, we leverage recent advancements in diffusion models. Specifically, we exploit random Gaussian noise to represent the diverse possibilities learnt from the target view data. We introduce a Geometry-guided Cross-view Condition (GCC) strategy to establish explicit geometric correspondences between satellite and street-view features. This enables us to resolve the geometry ambiguity in-

troduced by camera pose between image pairs, boosting the performance of cross-view image synthesis. Through extensive quantitative and qualitative analyses on three benchmark cross-view datasets, we demonstrate the superiority of our proposed geometry-guided cross-view condition over baseline methods, including recent state-of-the-art approaches in cross-view image synthesis. Our method generates images of higher quality, fidelity, and diversity than other state-of-the-art approaches.

1. Introduction

Ground-and-satellite cross-view image synthesis has attracted considerable attention recently due to its potential applications in virtual reality, simulations, cross-view image matching and data augmentation, *etc*. The task is to synthesize a target view image from a given viewpoint and a relative pose between the two views. The synthesized images are expected to not only exhibit a geometrically consistent scene structure between the views but also maintain high visual fidelity to real-world data.

The cross-view image synthesis is a remarkably challenging and inherently ill-posed learning task. This complexity arises primarily from the drastic viewpoint change, which results in minimal Field-of-View (FoV) overlap, severe occlusion, and large discrepancies in image contents and visual features. Preliminary works in cross-view synthesis mostly relied on conditional Generative Adversarial Networks [20]. Some of them focus on generating corresponding ground-view conditioned on a given satellite image patch, employing high-level semantics or contextual information for supervision [19, 24, 25, 42, 54]. Recent research [14, 22, 33] has further proven that incorporating 3D geometry into the learning process can significantly boost the quality of generated ground-view images. However, all these works formulate the task as a deterministic image-to-image translation, while the ground-and-satellite cross-view synthesis is inherently a probabilistic one-to-many problem.

Diffusion models have emerged as a powerful new family of deep generative models and have achieved state-of-the-art results in generative tasks, especially in image generation [3, 7, 40]. The recent Latent Diffusion models (LDM) [3] have enabled the probabilistic generation of high-quality images from any prompts, making it a preferable option to model the uncertainty in the ground-and-satellite cross-view synthesis task.

Most of the recent researches follow the path of Text-to-Image generation, utilizing superior power of vision-language models such as CLIP [23]. Zero123 [17] demonstrates a way to prepare an image condition with its camera pose information by concatenating image CLIP encoding and frequency embedded camera pose. Then using it as a conditioning representation for fine-tuning the pre-trained Stable Diffusion Model to learn posed CLIP embeddings. This conditioning method has demonstrated promising performance in the multi-view synthesis task at object level, but the model needs to implicitly learn the relationship between the conditioning image, pose, and the target image. We have discovered that image CLIP embedding is insufficient in generating cross-view images with fine-grained geometric accuracy and spatially alignment due to underlying ambiguous relationship between the ground-view image and camera pose, please see Sec. 4.4.1 for more details.

To address the ambiguity mentioned above, we propose Cross-view Diffusion, a conditioned cross-view synthesis framework developed upon LDM, as described in Fig. 2. Instead of using the widely accepted pretrained CLIP image encoder[17, 53], this paper leverages a Geometry-guided Cross-view condition that is derived from explicit 3D geometric projection of extracted image features. The proposed condition demonstrates capability in generating cross-view images of high quality and fidelity with fine-grained geometric and semantic control. Importantly, with the same framework, our method is able to handle both ground-to-satellite and satellite-to-ground image synthesis, where Grd2Sat is considered more challenging [24] due to the limited FoV of ground images and the presence of occlusion in ground views. The main contributions of this paper are summarized as follows:

- We present Geometry-guided Cross-view Diffusion, a conditional generative framework for cross-view image synthesis. Our approach showcases state-of-the-art performance in synthesizing images for both Sat2Grd and Grd2Sat tasks across multiple cross-view datasets.
- We propose a Geometry-guided Cross-view Conditioning approach, a novel 3D geometry-aware condition to guide the generation process of diffusion models. This conditioning approach eases the burden of diffusion models without implicitly learning the cross-view domain discrepancy and pose ambiguity from ground cameras, allowing our framework to generate geometrically and semantically consistent cross-view images.
- Our method is able to generate plausible target-view images with diversity from single input view, successfully modeling the one-to-many property of the ground-and-satellite cross-view image synthesis task.

2. Related Work

2.1. Ground-and-satellite image synthesis

Mapping ground and satellite images from one domain to another was first explored by Zhai *et al*. [54]. They proposed to learn a linear transformation matrix between satellite and street-view semantics. Later, Regmi and Borji [24, 25] demonstrated that the conditional GANs could effectively address the ground-and-satellite cross-view image

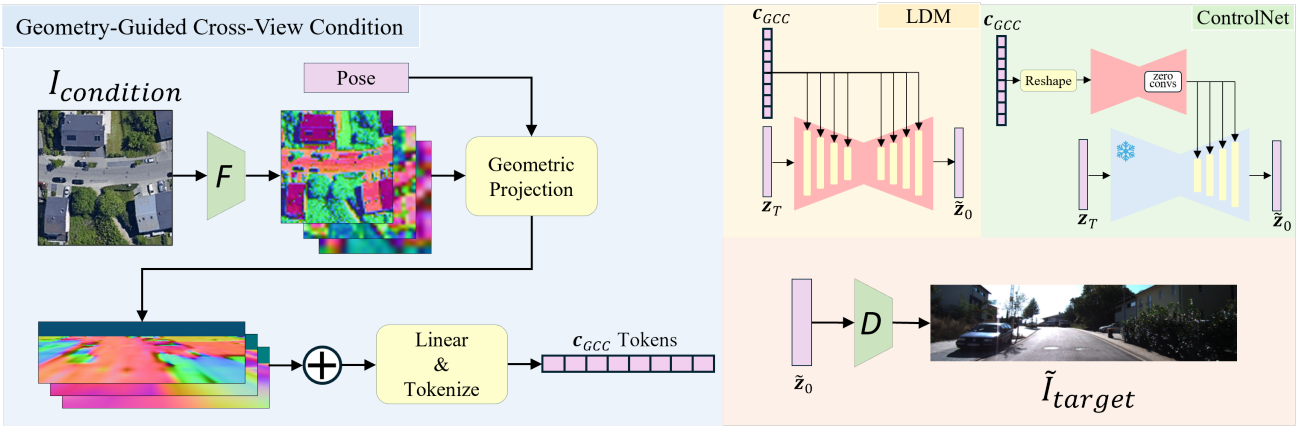


Figure 2. An overview of the proposed Cross-view Image Synthesis Pipeline. When provided with either a satellite image patch or a street-view image, the model employs a feature extractor $\mathcal{F}$ and our Geometry projection Module to construct our Geometry-guided Cross-view Conditions(GCC). The Latent Diffusion Pipeline learns to model cross-view data distribution from a Gaussian noise latent, under the guidance of our proposed GCC. The ControlNet module takes GCC as input and fine-tunes LoRA layers.

synthesis task, and adding an additional branch to the network for semantic map estimation could facilitate the view synthesis quality. After that, different powerful networks have been exploited for this task [42, 48, 61].

Recently, researchers explored how to combine the geometric correspondences between the views for the cross-view synthesis task [14, 19]. Lu *et al*. [19] proposed first estimating height and semantic maps from satellite images, which was then used to recover the ground structure and prepare a better condition for the ground-view image synthesis. Considering this method requires GT height and semantic maps of satellite images for training, Shi *et al*. [33] modeled geometric correspondences between the cross-view pixels in an end-to-end framework, eliminating the need for semantic and satellite map height supervision. The most recent work, Sat2Density [22], further improved the synthesis quality by modeling the transparency and illumination of sky regions.

However, all the above works formulate the task as a deterministic mapping. In the context of image synthesis, the relationship between the generative power of diffusion models and the representation of conditions is still an under-explored area of research [27]. The ground-and-satellite cross-view synthesis is inherently a one-to-many task due to the severe occlusions and illumination differences between the views. This paper resorts to the recently advanced diffusion models to address the probabilistic nature.

2.2. Cross-view image-based localization

Ground-to-satellite camera localization aims to determine a ground camera's location against a satellite map. The task was proposed initially for city-scale localization and formulated by cross-view image retrieval [1, 2, 8, 15, 16, 21, 26, 30–32, 35, 41, 43–47, 52, 57, 59, 60]. In this task, many works [26, 43] have demonstrated that the cross-view image synthesis objective is beneficial to improve the cross-view localization performance. Recently, researchers have extended the task to fine-grained pose refinement once the most similar satellite image has been retrieved for the query image [4, 13, 28, 29, 34, 36, 49–51, 58]. In this line of work, the cross-view feature synthesis has been extensively explored. Shi and Li [29] exploited satellite-to-ground feature synthesis because ground images have a larger resolution of scene contents and thus is sensitive to camera location change. Fervers *et al*. [4], Shi *et al*. [34] and OrienterNet [28] leveraged ground-to-satellite synthesis as registering synthesized overhead view feature map to reference satellite feature map leads to more efficient camera pose computation. Instead of cross-view feature synthesis, this paper addresses the problem of cross-view image synthesis. We expect this task can potentially facilitate cross-view localization performance by introducing more readily available cross-view image pairs for localization network training.

3. Methods

3.1. Problem Formulation

This paper focuses on synthesizing corresponding cross-view images conditioned on either a ground-view image or a satellite image, so we address the task as a conditional image generation problem. Given a satellite image patch $I_s \in \mathbb{R}^{H_s \times W_s \times 3}$ or a ground image $\mathbb{R}^{H_g \times W_g \times 3}$ along with a relative pose between the cross-view input-target images, we expect our model to learn to synthesize corresponding ground image $I'_g \in \mathbb{R}^{H_g \times W_g \times 3}$ or satellite image $I'_s \in \mathbb{R}^{H_s \times W_s \times 3}$, respectively, under the guidance of the conditioned input. The dimensions of the image prompt and desired generation target from the same domain are kept identical for simplicity.

We then follow LDM [27] to train our model in a learned image latent space, a perceptually equivalent space to the data space, but is more suitable for likelihood-generative models and more computationally efficient. We aim to reconstruct the latent $\mathbf{z}_0$ from a Gaussian noise sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ is a desired data point encoded by a learnt image encoder. Subsequently, the training objective of the denoising process is formulated as:

$$\mathcal{L}_{LDM} := \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{x}),\mathbf{c}_{GCC},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}_{GCC})\|_2^2\right], \tag{1}$$

which is guided by our proposed Geometry-guided Cross-view Condition $\mathbf{c}_{GCC}$. Overview of the framework is shown in Fig. 2, more details on LDM and its formulations will be included in the supplementary material.

3.2. Geometry-guided Cross-view Condition

In this paper, we address cross-view image synthesis as a conditional image generation task. We aim to explore image synthesis conditioned on multiple guidance from different domains: a correspondent cross-view image pair with the ground camera's relative pose to the satellite image.

To tackle the ambiguity caused by the relative camera pose between the two views, we propose Geometry-guided Cross-view Condition (GCC), a novel approach that effectively embeds camera pose information into our multi-level image features using our Cross-View Geometry Projection (CVGP) Module denoted as $\mathcal{P}$. The CVGP module projects the given viewpoint image feature representations to the target viewpoint according to the relative camera pose. The projected multi-level features are adopted as the diffusion model condition. This approach bridges the domain gap between the conditioning image and pose prompt and the desired distribution for target images.

3.2.1 Cross-View Geometry Projection

In this section, the geometric projection from satellite image to ground panoramic image is used as an example to illustrate the pixel mapping process. We begin by defining the geometric relationship involving the reference world coordinate system, ground camera pose, and the center of the reference satellite image. In this coordinate system, its origin is set to the reference satellite image center, the x axis aligns with latitude or the v_s direction, the y axis points downward into the image, the z axis is parallel to longitude or the u_s direction. Then, we can map any point $[x, y, z]^T$ in the world coordinate system to a satellite pixel coordinate with an orthogonal projection:

$$[u_s, v_s]^T = [\frac{z}{\gamma} + u_s^0, \frac{x}{\gamma} + v_s^0]^T, \tag{2}$$

where γ is the per-pixel real-world distance of the satellite feature map, and $[u_s^0, v_s^0]$ is the center of the satellite feature map. Given the location correspondence between the street-view camera position and the satellite image center, we are able to derive the pixel correspondences with cross-view geometry. We define a cylindrical image plane to represent an omnidirectional street-view image, with its pixels parameterized under a spherical coordinate system. For any satellite pixel $p_s = [u_s, v_s]^T$ that is visible from the ground view, we derive the pixel mapping to its projected pixel $p_g = [\theta, \phi]^T$ in the ground-view panorama with height z:

$$\theta = \begin{cases} \text{atan2}(\sqrt{(v_s - v_s^0)^2 + (u_s - u_s^0)^2}, \frac{-y}{\gamma}) & \text{if } y \neq 0 \\ \pi/2 & y = 0 \end{cases}$$
$$\phi = \begin{cases} \text{atan2}(v_s - v_s^0, u_s - u_s^0) & \text{if } u_s \neq u_s^0 \\ \pi/2 \cdot \text{sign}(v_s - v_s^0) & u_s = u_s^0 \end{cases} \tag{3}$$

Likewise, we can derive the mapping from a ground pixel to a satellite pixel and accommodate other camera projection models, such as a pinhole camera. Please refer to Appendix C for extensive details.

3.2.2 Multi-Level Projected Feature Aggregation

We utilize an arbitrary feature extraction backbone to extract visual features at multiple levels for preserving both high-level semantic information and of the input image at both coarse and fine levels. The geometric projection derived in the previous section only approximates the pixel correspondence, which might be prone to distortions and misalignment. Therefore, we project the extracted multi-level deep features rather than RGB pixels for robustness towards defects caused by our geometric assumptions, such as the ground camera height and ground plane Homography. Our extensive experiments also prove that using projected RGB images as the condition results in limited performance.

In the Grd2Sat image synthesis pipeline, we begin by extracting ground feature representation $\mathbf{F}_g \in \mathbb{R}^{H_g^l \times W_g^l \times C^l}$, where $l = 1, \ldots, L$ denotes the level of features. We then project these feature representations by our geometry projection module. Subsequently, we interpolate the feature maps at each level to a unified uv grid with bilinear interpolation, resulting in projected features $\mathbf{F}_{g2s}^l \in \mathbb{R}^{H_c \times W_c \times C^l}$. The final Geometry-guided Cross-view Condition is obtained by linearly mapping the aggregation of projected features from every level to the condition dimension:

$$\mathbf{c}_{g2s} = Linear(\mathcal{P}(F_g^1) \oplus \cdots \oplus \mathcal{P}(F_g^L)) \in \mathbb{R}^{H_c \times W_c \times C_c}. \tag{4}$$

where the total number of visual condition token equals $H_c \times W_c$, each token with condition dimension C_c.

The condition preparation procedure for the Sat2Grd image synthesis pipeline stays analogous, differing only in the height and width of the extracted features. The condition is then processed into tokens, containing multi-resolution contextual and textural information. Utilizing the cross attention mechanism in diffusion models, these tokens can be regarded as visual sentences, enforcing visual and spatial consistency in the generated imagery content.

4. Experiments and Results

4.1. Dataset

We evaluate the performance of our method and compare it with other SOTA cross-view image synthesis methods on several benchmark datasets, including the cross-view KITTI[5], CVUSA [54] and aligned CVACT [33] datasets.

The cross-view KITTI dataset is splitted into three subsets, one training set and two test sets. We train our model with Training set and to ablate the performance of various conditions on both Training and Test1 sets, which are collected from the same region. We use the left ground image and its corresponding satellite image as a cross-view image pair.

CVUSA and CVACT are both cross-view datasets with panoramic ground images. CVUSA contains 35,532 location aligned cross-view image pairs for training and 8,884 image pairs for testing. However, the street-view image in CVUSA dataset are cropped at the top and bottom by Zhai *et al*. [54], with unspecified portions. Aligned CVACT is the location aligned split proposed for cross-view image synthesis task, it contains 26,519 training pairs and 6,288 testing pairs, which is processed to remove the misalignment between cross-view image pairs. During training and testing, we approximate the street-view image in the CVUSA dataset as having a 90° vertical field of view (FoV) with the central horizontal line corresponding to the horizon, and consider CVACT ground images with 180° horizontal visualization.

4.2. Implementation Details

In our implementation, we train our Sat2Grd model with 256 $\times$ 256 resized satellite image (approximately 50x50 m^2 ground coverage) and Grd2Sat model with 128 $\times$ 512 street-view image as our condition view. The synthesized cross-view images are 128 $\times$ 512 street-view images and 256 $\times$ 256 satellite patches for a fair comparison with existing approaches [22, 24, 33]. We follow Shi et al [33] to approximate the height of the street-view camera as 1.65 meters for the KITTI dataset, and 2 meters for the CVUSA and CVACT datasets when we perform our geometric projections.

To clarify, the LDM model in our results refers to a diffusion model trained from scratch with our proposed GCC, and the ControlNet [55] model refers to a module fine-tuned based on pretrained Stable Diffusion 2.1 model[27] with proposed GCC. We implement our model based on Latent Diffusion Model's [3] architecture with feature dimension of 768, and our image latents are obtained with a pretrained VAE image encoder [11]. LDM models are trained with a batchsize of 48 on two NVIDIA GeForce RTX 3090 for 500 epoches for LDM and 200 epoches for ControlNet. We set $T = 50$ as DDIM [38] sampling steps when inferencing samples. For extracting proposed GCC, we adopt Swin Transformer V2 [18] as our feature extractor and we construct our geometry-guided cross-view guidance with feature output from block 1, 3 and 5 from Swinv2. Furthermore, we utilize pinhole camera projection model for KITTI data and spherical camera model for CVUSA and CVACT in our Cross-view Geometry Projection Module.

4.3. Evaluation Metrics

In this research, We adopt two pixel-wise similarity and two learned feature similarity as metrics for quantitative evaluation. The structure similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) for measuring the pixel-wise similarity between two images. We further use Learned Perceptual Image Patch Similarity (LPIPS) [56] to evaluate the feature similarity of generated and real images. In our ablation study, we adopt VGG backbone, and for fair comparison with other methods, we employ the pretrained AlexNet [12] and Squeeze [9] networks as backbones for the LPIPS evaluation, denoted as P_{alex} and P_{squeeze}, respectively. We also include FID [6] as a measure for the similarity between our generated images and the real images from our datasets, which is proven to be consistent with increasing disturbances and human judgment.

4.4. Ablation Study

4.4.1 Effectiveness of Geometry-Guided Cross-view Condition

In this section, we first conduct detailed experiments on all three datasets to validate the importance of geometry guidance in diffusion model based cross-view image synthesis. Experiments were implemented to test for following conditions: *Original Image, Projected Image, Projected Feature, and proposed Geometry-guided Cross-view Condition(GCC).* The Original Image condition is obtained by encoding the original image with CLIP [23], the Projected Image condition refers to geometric projected original RGB image, and the Projected Feature is obtained by projecting the last hidden state of the feature encoder.

Camera Alignment and Pose Ambiguity We present quantitative comparisons in Tab. 1 on KITTI dataset to justify the importance of geometry information in our condition design. This analysis is conducted within the **Grd2Sat**

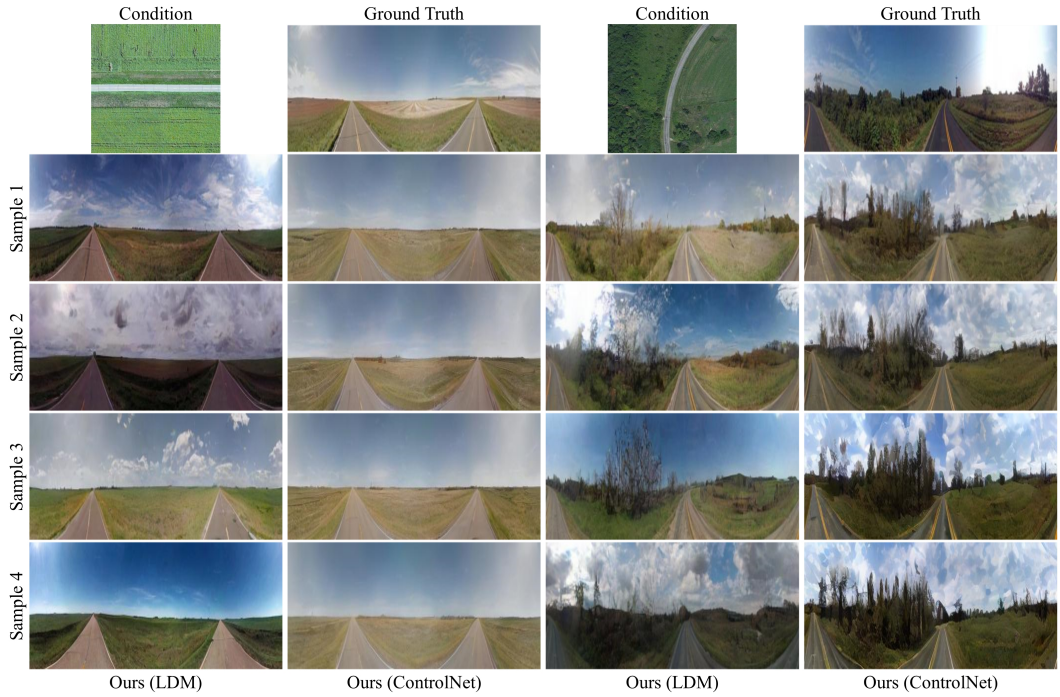


Figure 3. Ablation for sample generated by our LDM model and our ControlNet model given the same condition, on **Sat2Grd** task.

Table 1. Ablation study on significance of Geometric Guidance with the KITTI Dataset, under both Camera-aligned and North-aligned Setting.

		KITTI train			KITTI test1		
	Condition Prompt	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Camera-Aligned	Original Image	17.4063	0.3194	0.1968	**16.6424**	**0.2915**	**0.2226**
	Projected Image	15.0128	0.2030	0.3399	14.2573	0.1805	0.3903
	Projected feature	**17.9851**	**0.3511**	**0.1824**	16.4536	0.2851	0.2558
North-Aligned	Original Image	14.533	0.1677	0.3434	13.893	0.152	0.3894
	Projected Feature	**17.082**	**0.2860**	**0.1896**	**15.274**	**0.2131**	**0.3106**

Table 2. Comparison between our proposed GCC and other baseline conditions with Sat2Grd synthesis on CVUSA and CVACT datasets.

	CVUSA			CVACT		
Condition Prompt	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Original Image	12.078	0.3229	0.5488	13.46	0.3815	0.4728
Projected Feature	12.550	0.3363	0.4858	13.118	0.3355	0.4536
GCC (LDM)	14.032	**0.3589**	**0.4665**	15.286	0.4332	0.4592
GCC (ControlNet)	**14.274**	0.3420	0.5063	**15.609**	**0.4485**	**0.4136**

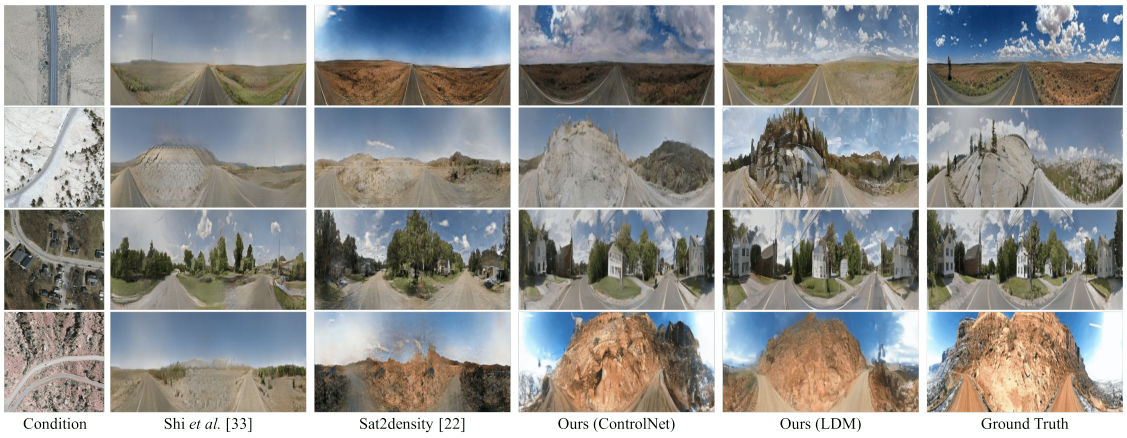
synthesis setup, where we evaluate the model's performance in generating Camera-aligned and North-aligned satellite images with various conditions. please see Appendix D for details about Camera-aligned and North-aligned setting.

As shown in the upper section of Tab. 1, when the satellite image is Camera-aligned, using original image as condition can achieve comparable performance as the projected feature. However, when the satellite image is north aligned, the geometric relationship between the ground camera and the satellite remains unresolved. This pose ambiguity might hinder the model to establish connection between the condition and the data distribution. The North-Aligned results exhibit a significant decline in performance when using the original image as a condition, whereas the projected feature demonstrates robustness against the introduction of pose ambiguity. Hence, it can be suggested that providing a geometrically aligned condition is able to enhance the model's ability to learn to associate the condition and the learned data distribution.

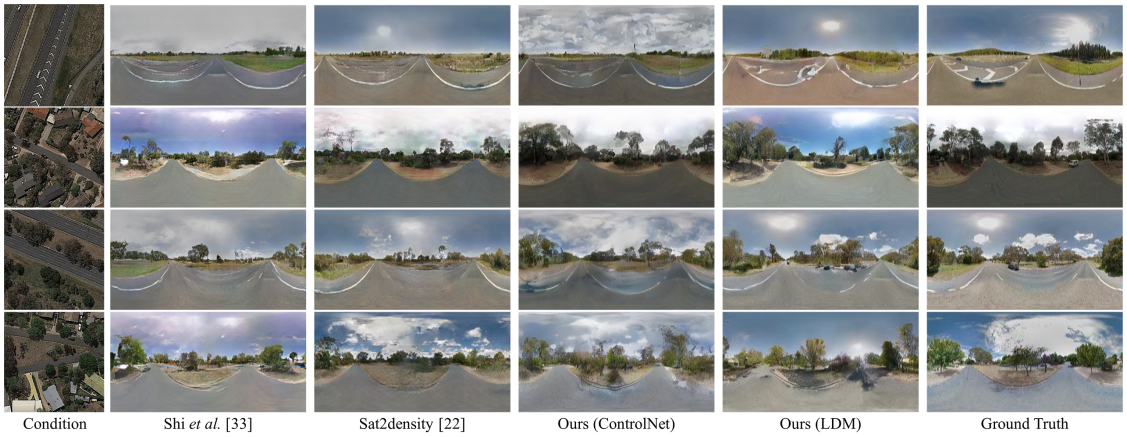
Capability of Conditions To further assess the effectiveness of conditions, we conduct a comparative analysis of their performance by training LDM and ControlNet models on the **Sat2Grd** task using CVUSA and CVACT datasets. From Tab. 2, the utilization of our proposed condition significantly improves all three metrics for both LDM and ControlNet models. Meanwhile, LDM is still capable of learning the hidden spatial information from its structure and generate images with consistent semantic content compared to the ground truth image. This validates our assumption that tokenized projected feature sequence is capable of building dense spatial correspondence with the input latent with the cross-attention mechanism, even without any positional hints.

4.4.2 Generative Diversity of LDM and ControlNet

The main reason we present results from both LDM and ControlNet is that the two models demonstrate different level of diversity in the **Sat2Grd** task, as shown in Fig. 3. For fair comparison, we disable classifier-free guidance scale for both models when inference. The LDM model trained from scratch is able to offer reasonable variations to the generated image, in terms of illumination, scene object and sky. While the ControlNet model generates images



(a) CVUSA Sat2Grd Visualization



(b) CVACT Sat2Grd Visualization

Figure 4. Example of generated images by different methods in **Sat2Grd** image synthesis task, on the CVACT (Aligned) and CVUSA datasets.

Table 3. Quantitative comparison with existing **Sat2Grd** image synthesis algorithms on the CVACT and CVUSA datasets.

	CVUSA					CVACT				
Method	PSNR↑	SSIM↑	P_{alex} ↓	P_{squeeze} ↓	FID ↓	PSNR↑	SSIM↑	P_{alex} ↓	P_{squeeze} ↓	FID ↓
Pix2Pix [10]	13.48	0.2946	0.5092	0.3902	-	14.38	0.3852	0.4654	0.3096	-
XFork [24]	13.68	0.2873	0.5144	0.4041	-	14.50	0.3710	0.4638	0.3262	-
Shi *et al*. [33]	13.77	0.3451	0.4639	0.3506	44.092	14.59	0.4272	0.4059	0.2708	49.401
Sat2Density [22]	13.78	0.3301	0.4504	**0.3365**	38.078	14.92	**0.4586**	0.3842	0.2573	38.029
Ours (LDM)	14.032	**0.3589**	0.4410	0.3414	17.501	15.286	0.4332	0.3915	0.2669	**21.638**
Ours (CtrlNet)	**14.274**	0.3420	**0.4345**	0.3397	**13.755**	**15.609**	0.4485	**0.3765**	**0.2550**	23.706

with less diversity but finer-grained similarity to the target image. Additionally, we noticed that both models seem to demonstrate limited generative diversity on the **Grd2Sat** task. We attribute this phenomenon to the property of data: image contents captured by satellite images are not as diverse as the ground view images, especially the sky itself can offer variety of illumination. Furthermore, geometric projection from ground-view to satellite view is ill-posed, ground features can only provide limited information that aligns with the satellite image, hence it is a harder task to learn compared to **Sat2Grd**. Please see Appendix B.2 and Appendix E for further explanation and examples for the two models.

4.5. Comparison with existing methods

In the **Sat2Grd** view synthesis task, we compare our methods with Pix2Pix [10], XFork [24], Shi *et al*. [33] and Sat2Density [22]. As for **Grd2Sat** view synthesis, we compare our methods with Pix2Pix [10] and XFork [24]. Pix2Pix and XFork are classic conditional GAN-based models designed for learning image-to-image translation. However, they do not incorporate the relationship

Table 4. Quantitative comparison with existing **Grd2Sat** image synthesis algorithms on the CVACT and CVUSA datasets.

Method	CVUSA					CVACT				
	PSNR↑	SSIM↑	P_{alex} ↓	P_{squeeze} ↓	FID ↓	PSNR↑	SSIM↑	P_{alex} ↓	P_{squeeze} ↓	FID ↓
Pix2Pix [10]	11.33	0.1229	0.5490	0.3931	162.505	11.60	0.0462	0.6692	0.3462	70.168
XFork[24]	10.85	0.1037	0.5908	0.4301	156.252	11.63	0.0656	0.6811	0.3716	184.283
Ours (LDM)	12.838	0.1751	0.5093	0.4045	70.125	14.537	**0.1384**	0.4998	0.2561	53.646
Ours (CtrlNet)	**14.070**	**0.2271**	**0.4829**	**0.2866**	**53.080**	**14.598**	0.1375	**0.4849**	**0.2455**	**33.560**

Figure 5. Example of generated images by different methods in **Grd2Sat** image synthesis task, on the CVACT (Aligned) dataset.

between cross-view images with 3D geometry. Shi et al and Sat2Density are both geometry-guided synthesis model. The former represents 3D geometry using a depth probability multiplane image, while the latter introduces a framework to learn a density field representation from cross-view image pairs and synthesis ground-view panoramas based on learned 3D cross-view geometry. A notable advantage of our method is that we do not require additional input such as segmentation maps and accurate height information to construct our condition.

4.5.1 Quantitative Comparison

We report the average metrics of **10** generated samples per image condition. As presented in Tab. 3, it is evident that our models outperform other methods on low-level and perceptual similarity measures such as SSIM, P_{alex}, P_{squeeze} and FID in the **Sat2Grd** task. The relatively low PSNR is also in line with our expectations due to the non-deterministic nature of diffusion models. Instead, we anticipate our model to demonstrate more diversity in the generated samples, while maintaining promising structural integrity and high level alignment. It is worth noting that the Sat2Density-oracle model generates ground-view images with sky histogram from the ground truth image. Given that the sky region typically occupies nearly half of the ground-view images, possessing the ground truth sky histogram and illumination hints grants it a significant advantage in terms of pixel similarity. Therefore, simply comparing our results against theirs solely based on PSNR would be unfair. In the supplementary material, we also provide evaluation results without sky region against Sat2Density.

The quantitative results of **Grd2Sat** are displayed in Tab. 4. Our models also achieve the best performance on all metrics, almost doubled the SSIM score of the existing researches. In both tasks, we show a notable improvement on FID score. This proves that our model is capable of learning the probabilistic distribution of the scene objects in the datasets, instead of learning one-to-one relationship to generate image only matches its ground truth target. As a result, we can generate image with high fidelity and close to real satellite and street view images.

4.5.2 Qualitative Comparison

Fig. 4 displays the **Sat2Grd** synthesis results, Shi *et al*. [33] and Sat2Density [22] can generate reliable 3D geometry like road direction, while our method can clearly display road lines and better predict invisible side facade and obstacle. The last three rows in Fig. 4a shows that our method can successfully synthesize building and landscape in the scene, while other two methods failed. In CVACT samples displayed in Fig. 4b, we show that our geometry and spatial alignment is more accurate.

Fig. 5 displays generated images on **Grd2Sat**, we find that Pix2Pix [10], XFork [24] is incapable of synthesizing reasonable satellite images' structure, but our method generates consistent geometry based on the ground view, like the road intersections, shape and direction of the roads. In addition, our model is able to provide reasonable prediction for the unseen region, to generate style-consistent building and plants along the street road.

5. Conclusion

This paper has presented a novel approach to cross-view image synthesis, addressing the inherent challenges of one-to-many mapping and uncertainty modeling. By leveraging the probabilistic diffusion models and establishing explicit geometric correspondences between views, we have demonstrated significant improvements in view synthesis quality in both ground-to-satellite and satellite-to-ground synthesis across various datasets. Moving forward, extending our methodology to incorporate additional modalities, such as text, depth information or learning across multiple datasets, could broaden its applicability and enhance its capabilities.

References

[1] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[2] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–17, 2015. 3

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 5, 1

[4] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21621–21631, 2023. 3

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5, 2

[6] 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, 2018. 5

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 1

[8] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[9] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 5

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 7, 8

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 5

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[13] Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17225–17234, 2023. 3

[14] 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*, pages 12436–12445, 2021. 2, 3

[15] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 3

[16] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2

[18] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. 5

[19] 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*, pages 859–867, 2020. 2, 3

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[21] Arsalan Mousavian and Jana Kosecka. Semantic image based geolocation given a map. *arXiv preprint arXiv:1609.00278*, 2016. 3

[22] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. 2023. 2, 3, 5, 7, 8

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 5

[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, 5, 7, 8

[25] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. page 102788. Elsevier, 2019. 2

[26] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3, 4, 5, 2

[28] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21632–21642, 2023. 3

[29] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17010–17020, 2022. 3

[30] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems*, pages 10090–10100, 2019. 3

[31] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[32] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *AAAI*, pages 11990–11997, 2020. 3

[33] 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. 2, 3, 5, 7, 8

[34] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21516–21526, 2023. 3

[35] Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. Cvlnet: Cross-view semantic correspondence learning for video-based camera localization. In *Computer Vision–ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part I*, pages 123–141. Springer, 2023. 3

[36] Yujiao Shi, Hongdong Li, Akhil Perincherry, and Ankit Vora. Weakly-supervised camera localization by ground-to-satellite image registration. In *European Conference on Computer Vision*, pages 39–57. Springer, 2025. 3

[37] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 1

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 5, 1

[39] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. 1

[40] 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. 2

[41] Bin Sun, Chen Chen, Yingying Zhu, and Jianmin Jiang. Geocapsnet: Ground to aerial view image geo-localization using capsule network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 742–747. IEEE, 2019. 3

[42] 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*, pages 2417–2426, 2019. 2, 3

[43] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. *CVPR*, 2021. 3

[44] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 494–509. Springer, 2016.

[45] Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 440–456. Springer, 2022.

[46] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–78, 2015.

[47] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 3

[48] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Jianjun Qian, Nicu Sebe, Yan Yan, and Qinghua Zhang. Cross-view panorama image synthesis with progressive attention gans. *Pattern Recognition*, 131:108884, 2022. 3

[49] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision*, pages 90–106. Springer, 2022. 3

[50] Zimin Xia, Olaf Booij, and Julian FP Kooij. Convolutional cross-view pose estimation. *arXiv preprint arXiv:2303.05915*, 2023.

[51] Zimin Xia, Yujiao Shi, Hongdong Li, and Julian FP Kooij. Adapting fine-grained cross-view localization to areas without fine ground truth. In *European Conference on Computer Vision*, pages 397–415. Springer, 2025. 3

[52] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021. 3

[53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

[54] 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*, pages 867–875, 2017. 2, 5

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 5, 2

[56] 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. 5

[57] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Cross-view image sequence geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2914–2923, 2023. 3

[58] Yanhao Zhang, Yujiao Shi, Shan Wang, Ankit Vora, Akhil Perincherry, Yongbo Chen, and Hongdong Li. Increasing slam pose accuracy by ground-to-satellite image registration. *arXiv preprint arXiv:2404.09169*, 2024. 3

[59] Sijie Zhu, Taojiannan Yang, and Chen Chen. Revisiting street-to-aerial view image geo-localization and orientation estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 756–765, 2021. 3

[60] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1162–1171, 2022. 3

[61] Yingying Zhu, Shihai Chen, Xiufan Lu, and Jianyong Chen. Cross-view image synthesis from a single image with progressive parallel gan. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

Geometry-Guided Cross-View Diffusion for One-to-Many Cross-View Image Synthesis

Supplementary Material

A. Overview

In this supplementary material, we provide the following relevant details that could not be included in the main paper:

1. More details on LDM and ControlNet Implementation
2. Additional details of the Geometry Projection Module.
3. Extended explanation of North-Aligned and Camera-Aligned setting of Ablation Study.
4. Extended Ablation Results
5. Additional Quantitative and Qualitative Results.
6. Visualization of Failure Cases

B. Additional Details of LDM and ControlNet implementation on Cross-view Diffusion

B.1. Diffusion Models

Preliminary. Diffusion models [7, 37, 38] are a class of latent variable models that have been proven to be superior to GANs in both unconditional and conditional image synthesis tasks [3]. It is capable of learning a data distribution from an isotropic Gaussian distribution by reversing a diffusion process.

Consider a forward diffusion process fixed to a Markov Chain that gradually adds Gaussian noise for a large number of timesteps T. The noising operator at each timestep $t \in \{1, \dots, T\}$ is defined as

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \quad (5)$$

By which we can compute the approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$ from $\mathbf{x}_0$ in the interested data distribution according to a variance schedule $\beta_1, \dots, \beta_T$ [7].

The reverse process is defined as a Markov Chain that performs sampling from $\mathbf{x}_T$ to $\mathbf{x}_0$. With each denoising step being expressed as a learned Gaussian transition parametrized by θ to approximate intractable true denoising distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \quad (6)$$

Ho et al. [7] observe that the mean $\mu_\theta(\mathbf{x}_t, t)$ of the denoising model can be represented by a noise estimator network $\epsilon_\theta(\mathbf{x}_t, t)$ to predict ϵ from $\mathbf{x}_t$, then sample $\mathbf{x}_{t-1}$:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}, \quad (7)$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\bar{\alpha} = \prod_{s=1}^{t}(1-\beta_s)$.

Training of the denoiser network ϵ_θ is performed with denoising score matching over multiple noise scales indexed by t [39]:

$$\mathcal{L}_{DM} := \mathbb{E}_{\mathbf{x},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\lambda_t\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right] \quad (8)$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ and $\lambda_t = \frac{\beta_t^2}{2\sigma_t^2(1-\beta_t)(1-\bar{\alpha}_t)}$, practically setting $\lambda_t = 1$ for improved sample quality [7].

B.2. Difference between ControlNet and LDM models

The original intention of implementing the ControlNet model with the pretrained SD model is to utilize its strong visual prior learned from millions of images. However, Stable Diffusion is essentially a text-to-image model. When there is no appropriate language prompt, our condition inputs serve as constraints during the image generation process, which might limit its generative capability.

The main reason for us to report results from both ControlNet and LDM models is that we observe different performance that cannot be reflected solely on the quantitative metrics. As reported in Fig. 3, the LDM model trained from scratch can generate images with much diverse variation than the ControlNet model, in illumination, scene objects such as trees, boulders and buildings. Although the LDM model under-performs the ControlNet model in terms of quantitative metrics, its performances and synthesized images aligns better with our motivation to generate diverse image samples with the same condition.

LDM Implementation. Incorporating our proposed Geometry-Guided Cross-View Condition, our conditional denoising step can be expressed as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, c_{GCC}) := \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, c_{GCC}), \Sigma_\theta(\mathbf{z}_t, t, c_{GCC})). \quad (9)$$

Due to the computation resource limitation, our implementation deviates from configuration of the original Stable Diffusion model. We maintain the four blocks architecture of the LDM U-Net, but changed each block out channel size to $[240, 480, 960, 960]$, and also decreased the cross attention feature dimension from 1024 to 768.

ControlNet Implementation. As mentioned in Section 4 in the main paper, we have implemented a ControlNet

[55] version of our Cross-view diffusion pipeline for the effectiveness of our proposed Geometry-guided Cross-view Condition. Varying from the visual token sequence in the LDM [3] version, we pixel-wisely align our condition with the encoded image latent and input it to the ControlNet module by reshaping the input tensor (see Fig. 2). The ControlNet module is a trainable copy of the encoder section of the LDM UNet, connected to the decoder section by zero convolution layers, whereas the LDM parameters are frozen.

The pipeline is built upon pretrained Stable Diffusion 2.1 model [27], where the prompt input to the LDM Model should be text embedding. During the training of the ControlNet Module, we set the text prompt to be an empty string to assure our generation results are unaffected by the text conditioning. In the future, we might explore the effect of combining both ControlNet and text conditions.

C. Additional details of the Geometry Projection Module

Geometric Projection Derivation for Ground Camera with Pin-hole Model In this paper, we consider the 3-DoF (Degree of Freedom) ground camera pose for the KITTI [5] dataset, *i.e.*, the 1-DoF azimuth angle $\phi \in [-\pi, \pi]$ and 2-DoF translation along the latitude and longitude directions. Let $\mathbf{R} = \begin{pmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{pmatrix}$ and $\mathbf{t} = [t_x, 0, t_z]^T$ be the relative rotation and translation from real ground camera coordinate system to the world coordinate system and $\mathbf{K}$ be the ground camera intrinsics.

The back-projection from a pixel on a pin-hole camera image plane to the world coordinate system can be expressed as

$$[x, y, z]^T = w\mathbf{R}\mathbf{K}^{-1}[u_g, v_g, 1]^T + \mathbf{R}\mathbf{t} \tag{10}$$

where w is a scare factor.

By combining Eq. (2) from Sec. 3.2.1 and Eq. (10) above, we can derive the mapping from a ground-view pixel (u_g, v_g) to a satellite pixel (u_s, v_s) as

$$\begin{bmatrix} u_s \\ v_s \\ z \end{bmatrix} = \begin{bmatrix} \frac{1}{\gamma} & 0 & 0 \\ 0 & \frac{1}{\gamma} & 0 \\ 0 & 0 & 1 \end{bmatrix} \left(w\mathbf{R}\mathbf{K}^{-1} \begin{bmatrix} u_g \\ v_g \\ 1 \end{bmatrix} + \mathbf{R}\mathbf{t} \right) + \begin{bmatrix} u_s^0 \\ v_s^0 \\ 0 \end{bmatrix}. \tag{11}$$

The above projection is defined on ground plane homography, w is therefore computed based on the assumption of fixed camera height y_c. Similarly, we can derive the mapping from an satellite pixel to a ground image pixel

$$\begin{bmatrix} u_g \\ v_g \end{bmatrix} = \begin{bmatrix} f_x \frac{[(v_s - v_s^0) + t_x]\cos(-\phi) - [(u_s - u_s^0) + t_z]\sin(-\phi)}{[(v_s - v_s^0) + t_x]\sin(-\phi) + [(u_s - u_s^0) + t_z]\cos(-\phi)} \\ f_y \frac{h}{\gamma\left[[(v_s - v_s^0) + t_x]\sin(-\phi) + [(u_s - u^0) + t_z]\cos(-\phi)\right]} \end{bmatrix} + \begin{bmatrix} u_g^0 \\ v_g^0 \end{bmatrix}, \tag{12}$$

where f_x and f_y denote the ground camera focal length along u and v directions, respectively, h is the height of pixel (u_s, v_s) above the ground plane.

D. Extended Explanation of North-Aligned and Camera-Aligned setting

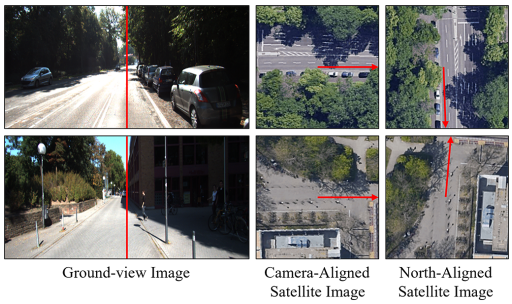


Figure 6. Example of Camera-Aligned and North-Aligned samples, the red arrows in the satellite views indicate the orientation of the ground camera.

As mentioned in Sec. 4.4 in the main paper, we presented ablation study results for camera-aligned and north-aligned setup on the KITTI dataset. As illustrated in Fig. 6, under the Camera-Aligned setting, the orientation of the ground-view image is always aligned in the same direction on the satellite view. When the satellite images are North-aligned, the orientation relationship between the satellite images and the ground-view image changes between pairs, which yields pose ambiguity between the cross-view image pairs that hinders the models' learning ability as reported in the Tab. 1 of the main paper. However, our experiment show that the model with projected feature condition suffers less performance drop under the North-aligned setting comparing to the image condition, which can effectively mitigate the influence of pose ambiguity.

E. Further Ablation results on the Generative Ability of Models

In Fig. 7, we show the qualitative ablation on the **Grd2Sat** task with generated samples from both LDM and ControlNet Models. As stated in the main paper, the generative ability for the **Grd2Sat** is limited by the variability of the data itself, therefore, we do not see much diversity in the generated samples compared to the samples from the **Sat2Grd** task.

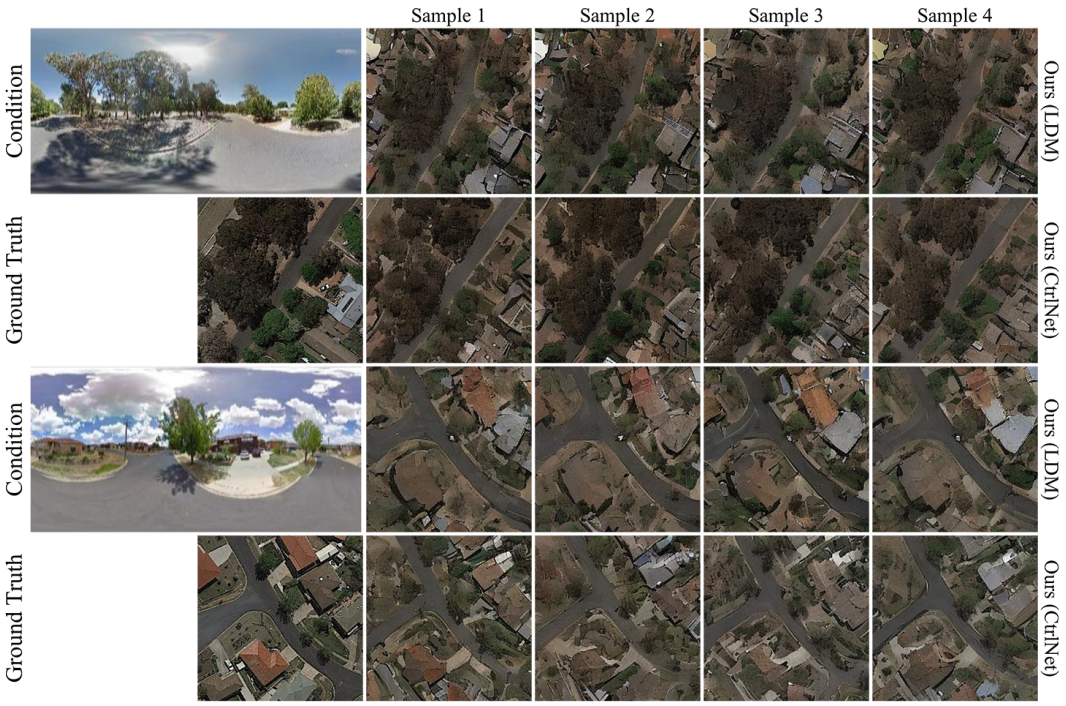


Figure 7. Ablation for sample generated by our LDM model and our ControlNet model given the same condition, on **Grd2Sat** task.

F. Additional Qualitative and Quantitative Results

Table 5. Overall Evaluation without sky region, on CVUSA, best in **bold**

Method	PSNR↑	SSIM↑	$P_{\text{alex}}\downarrow$	$P_{\text{squeeze}}\downarrow$
Sat2Density	14.528	0.2389	0.3958	0.3084
Ours(LDM)	14.791	**0.2908**	0.3867	**0.3074**
Ours(CtrlNet)	**14.879**	0.2725	**0.3861**	0.3090

In Fig. 8, we include qualitative comparisons of **Grd2Sat** results with existing methods on the CVUSA dataset. In Tab. 5, we conduct another evaluation with the sky regions excluded, evaluating only the shared region between the ground-view and satellite-view on the ground-level. Our results outperform Sat2Density in all metrics, showing that we are able to generate more geometrically and semantically aligned images with diversity.

G. Visualization of Failure Cases

In Fig. 9, we show some typical failure cases from **Grd2Sat**, on both CVUSA and CVACT datasets. The first two rows are samples from CVUSA, and last two rows are samples from CVACT.

In the first two rows, samples generated by our LDM model failed to reconstruct the true street structure, this might due to the model failed to pick up structural information from the given condition. As summarized in the main paper, our ControlNet version generally outperforms our LDM version in the **Grd2Sat** task, this might due to the stronger supervision from features that are pixel-aligned with the image latent. The samples generated in the third row failed to recover the shape of the round building, where the building shape can not be recognized simply by projecting the ground-view panorama. In the fourth row, the samples failed to generate the correct road structure at end of the road and also the car park behind the pedestrian walkway due to limited range of sight and occlusion in the ground-view.

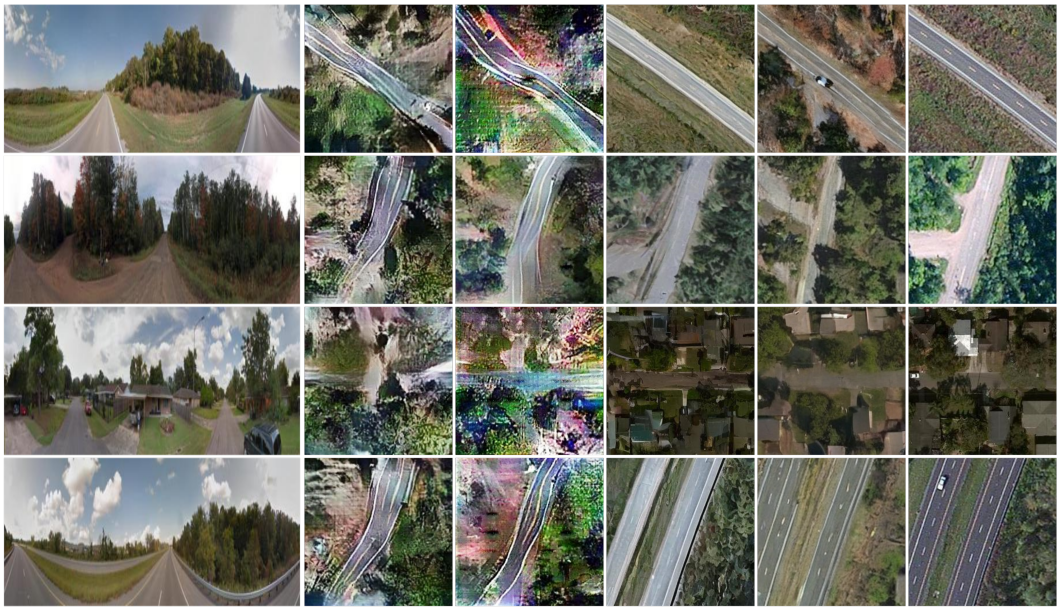


Figure 8. Qualitative comparisons of our results on the **Grd2Sat** task, on CVUSA dataset.

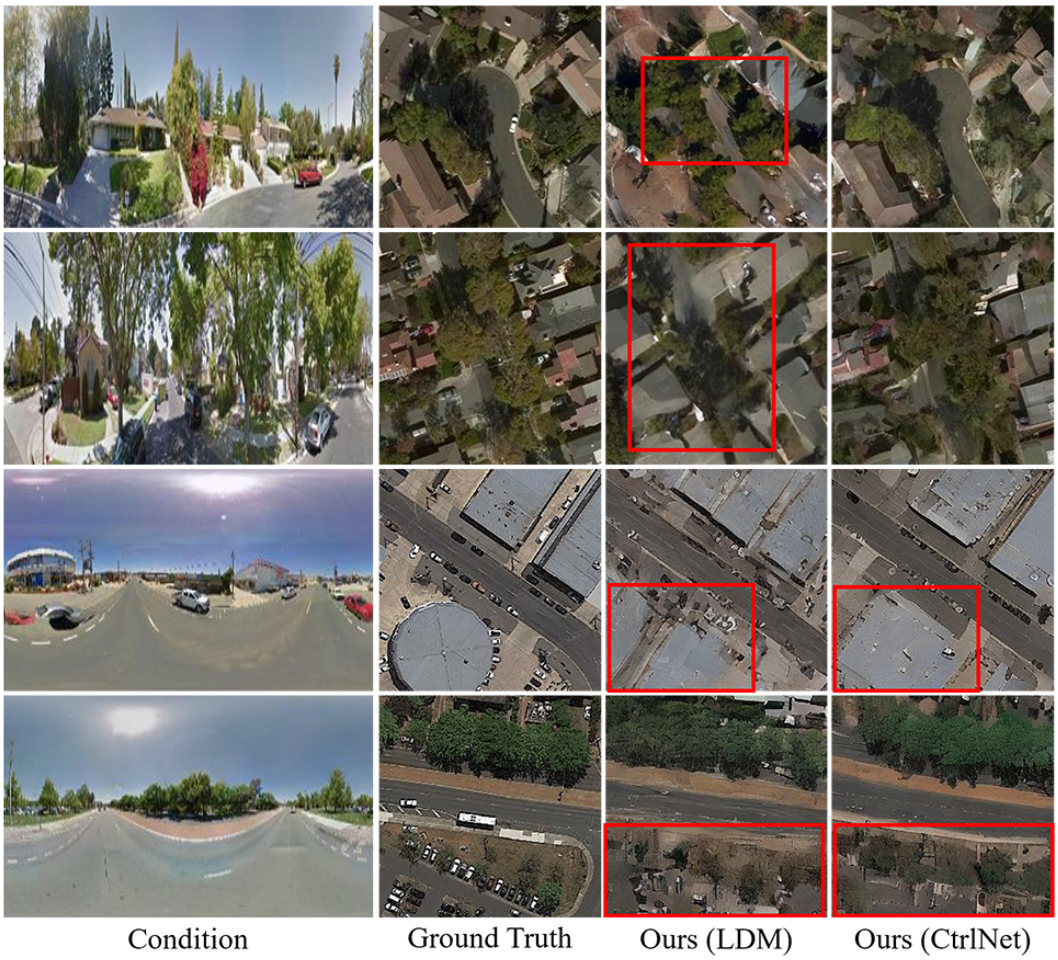


Figure 9. Some Failure cases on **Grd2Sat** task, on both CVUSA and CVACT datasets. We mainly visualize failure cases in **Grd2Sat**, as it is a much challenging task to learn and recover geometric and textural information by geometric projected feature alone, due to presence of limited range of sight (row 4), occlusion (row 2 and 4) and shape ambiguity (row 1 and 3).