

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, 36, 37] 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 β_1, \dots, β_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 $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 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 [38]:

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

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ 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].

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 [52] 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

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

where w is a scale factor.

By combining Eq. (2) from Sec. 3.3 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

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

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

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

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.

F. Additional Qualitative and Quantitative Results

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.

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

Method	PSNR \uparrow	SSIM \uparrow	$P_{alex} \downarrow$	$P_{squeeze} \downarrow$
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

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.

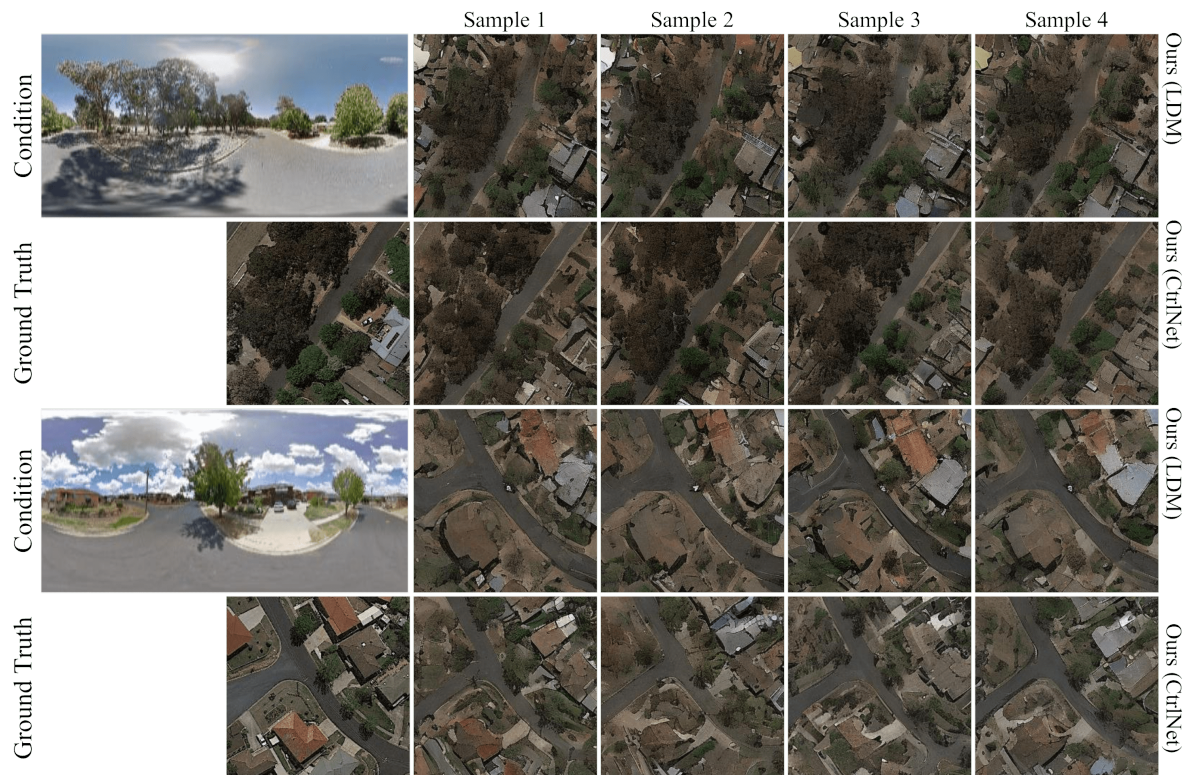


Figure 7. Ablation for sample generated by our LDM model and our ControlNet model given the same condition, on **Grd2Sat** task.

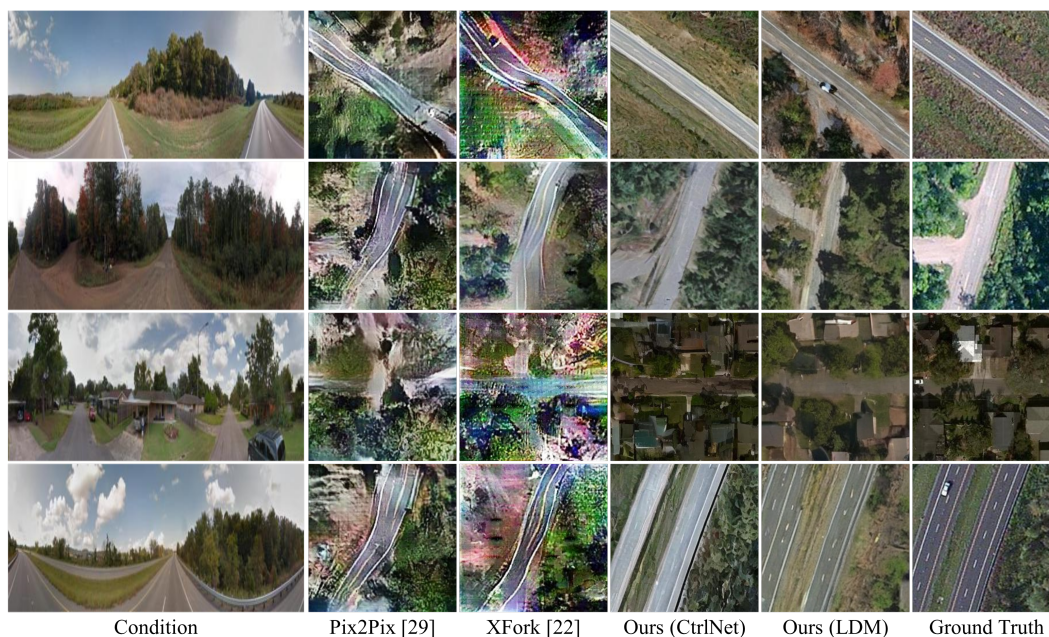


Figure 8. Qualitative comparisons of our results on the **Grd2Sat** task, on CVUSA dataset.

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 infor-

mation 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

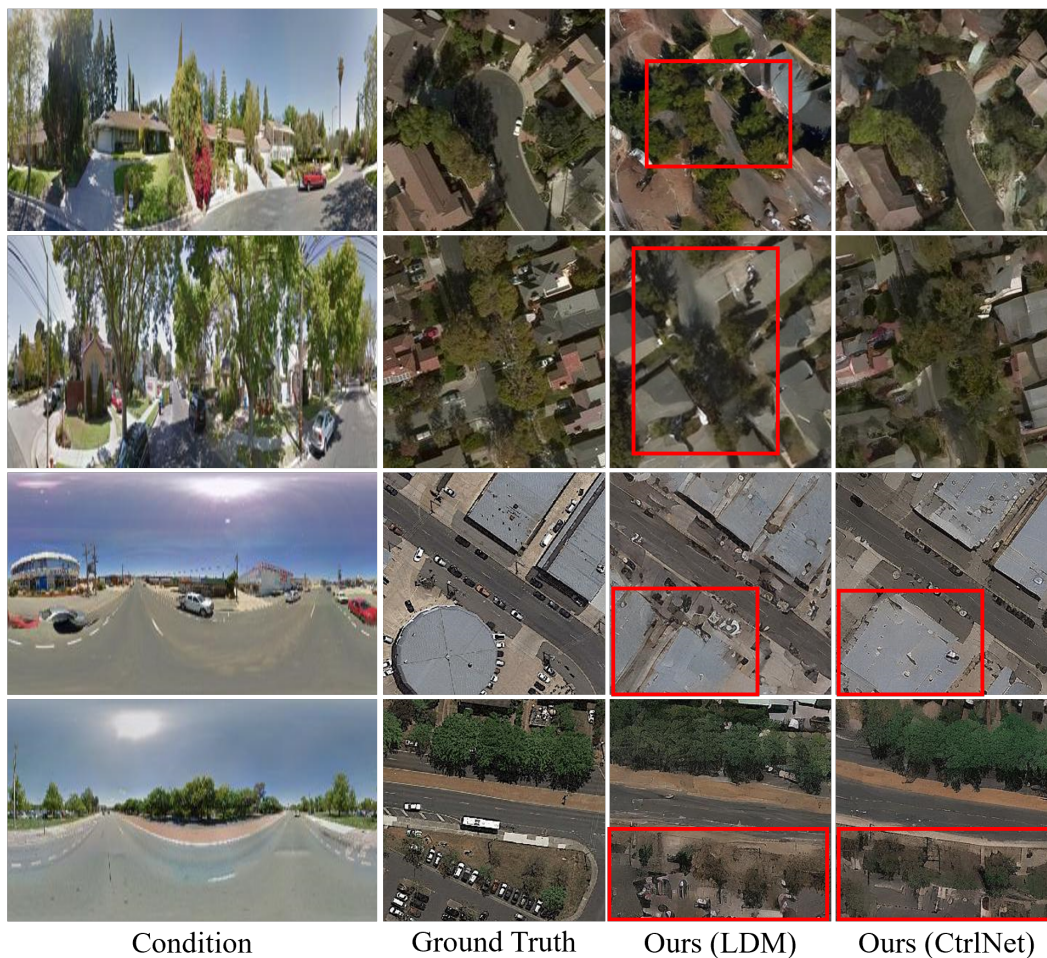


Figure 9. Some Failure cases on **Grd2Sat** task, on both CVUSA and CVACT datasets. We mainly visualize failure cases in **Grd2Sat**, as it is a much challenging task to learn and recover geometric and textural information by geometric projected feature alone, due to presence of limited range of sight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

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.