Geospecific View Generation - Geometry-Context Aware High-resolution Ground View Inference from Satellite Views

Ningli Xu and Rongjun Qin*

The Ohio State University, Columbus, OH 43210, USA
{xu.3961,qin.324}@osu.edu

Abstract. Predicting realistic ground views from satellite imagery in urban scenes is a challenging task due to the significant view gaps between satellite and ground-view images. We propose a novel pipeline to tackle this challenge, by generating geospecifc views that maximally respect the weak geometry and texture from multi-view satellite images. Different from existing approaches that hallucinate images from cues such as partial semantics or geometry from overhead satellite images, our method directly predicts ground-view images at geolocation by using a comprehensive set of information from the satellite image, resulting in ground-level images with a resolution boost at a factor of ten or more. We leverage a novel building refinement method to reduce geometric distortions in satellite data at ground level, which ensures the creation of accurate conditions for view synthesis using diffusion networks. Moreover, we proposed a novel geospecific prior, which prompts distribution learning of diffusion models to respect image samples that are closer to the geolocation of the predicted images. We demonstrate our pipeline is the first to generate close-to-real and geospecific ground views merely based on satellite images. Code and dataset are available at https://gdaosu.github.io/geocontext/.

Keywords: Cross-view synthesis · Conditional image generation · Cross-view geo-localization

1 Introduction

The growing availability of satellites offers the opportunity to capture images in every corner of the world. Directly predicting ground-view images from these images, referred to as the cross-view synthesis problem, can benefit numerous applications, such as 3D realistic gaming [18], and city-scale scene synthesis [22, 44].

The primary challenges lie in significant disparities in viewing directions and resolutions across satellite and ground-level domains. Firstly, the difference in

* corresponding author

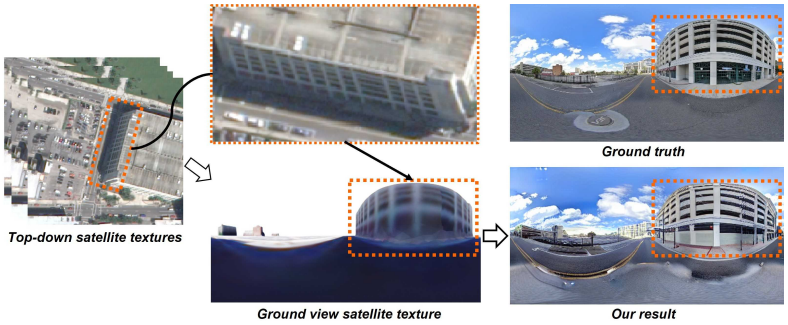


Fig. 1: Example of our synthesized geospecific views. Instead of conditioning on semantics [22,29,31], ours utilizes ground-view satellite texture which provides high-frequency structural and color information. The predicted result not only shows photorealistic quality but also accurately reflects the number of stories of the garage (marked as orange rectangles).

viewing angles makes the transformation from one view to the other very difficult and sensitive to noises. In urban areas, satellite images may capture subtle details of building facades. Transforming the visible facades to ground-view domains becomes highly sensitive to localization errors of building corners. Secondly, the low resolution of satellite images makes the extraction of useful information for ground-view synthesis difficult. The resolution of commercial satellite imagery is usually 0.3m/pixel, whereas the resolution of Google street-view imagery is much higher than satellite imagery, with around 3cm/pixel [13,45]. Bridging this nearly 10× resolution difference remains a challenge, which cannot be simply addressed by super-resolution techniques. Finally, due to the diffuse reflections of clouds, the color distortions between satellite imagery and ground-view images are significant.

Existing approaches [19,22,29,31,41] to address these challenges mainly seek for solutions that hallucinate views that reflect possible looks on the ground, which lack ground fidelity. They often adopt a black-box methodology or rely on auxiliary information. In tackling the disparity in viewing directions, [29,41] proposed end-to-end networks that directly learned the mapping relation between top-down satellite and ground-view images. As a result, the synthesized results lack photorealism and consistency in building facade regions due to substantial domain differences. [19, 22] bridges the viewing direction difference by leveraging known accurate geometry. They proposed a 2D-3D-2D projection method that first projects the top-down satellite texture into 3D space via orthographic projection and then projects the 3D satellite texture into ground-view 2D space by panoramic projection. Nonetheless, orthographic projection is a compromised solution that sacrifices facade information. To address resolution differences, the majority of existing works [19, 22, 29, 31] employ cGAN-based methods [4, 14], conditioning on ground-view semantics. However, this conditioning sacrifices tex-

ture information, making the synthesized results often deviate from the ground truth images.

Instead, our goal is to achieve ground view synthesis with not only photorealism but also maximal ground fidelity, meaning that the generated view will be geospecific, reflecting the actual looks at its geolocation. Our approach addresses the existing challenges for ground-view generation with a mathematically more accurate approach, avoiding any compromise of satellite texture and geometry information. Specifically, we proposed a novel cross-view synthesis approach with full utilization of satellite texture information including the visible building facades, as shown in Fig. 1. The projection from the top-down satellite to ground level is performed in the 2D-3D-2D way similar to [19, 22] while the difference is the utilization of an accurate satellite camera projection model (rational polynomial camera, RPC [37]). The noises of the satellite geometry usually cause the distortions of projected texture. We developed a texture-friendly geometry refinement method to minimize distortions of the projected satellite texture. Additionally, we present a geospecific prior approach to improve the training efficiency and synthesis quality. Experimental results demonstrate that all baseline methods utilizing such textures exhibit superior synthesis quality compared to those relying on semantics. Our synthesized results not only excel in all perceptual metrics but also accurately capture building facade layouts. We summarize the main contributions of this work as:

- The introduction of a texture-guided cross-view synthesis approach, which generates layout-preserving ground-view images conditioning building facade information.
- The development of a texture-friendly geometry refinement method allowing the utilization of subtle building facade details as the condition for cross-view synthesis.
- Through rigorous experiments, we demonstrate our method outperforms SOTA methods at various metrics including semantic resemblance, edge, and perception similarity.

2 Related Work

Cross-view synthesis focuses on the novel view synthesis of objects or scenes from a completely different view. A typical task is to synthesize ground views given top-down view satellite images. Its main challenges are the huge viewport and domain difference. Existing works bridge such differences by using the top-down view or extracting high-level features from the top-down view as the condition for the ground-view synthesis. [29, 41] used conditional GANs [14] predicting both the ground-view and corresponding semantics conditioning on top-down view satellite image. The large view-port difference makes such methods difficult to converge. Instead, [18,19,22,26] perform viewport transformation based on predicted geometry, where they estimated the height maps from top-down views assuming the orthographic projection. As an approximation projection, it can preserve the roof and ground information while ignoring the building

facade information. Our method developed a robust way of transforming such information to ground views while preserving the geometric consistency, which further served as the synthesis condition.

Conditional image generation focuses on learning a parametric mapping between source condition domains and target image domains. Example conditions include text descriptions to generate corresponding images [28, 33], broken images to fill the missing parts [23, 25, 33], street-view semantics to generate the corresponding images [14, 46, 49]. Among these works, conditional GANs [4, 14] are widely used as the backbone for conditional image generation while they suffer from slow and unstable convergence during the training process and require a large volume of paired data. Recently, diffusion models [10, 33, 38] have proven exceptional in image generation tasks, which iteratively denoises the Gaussian noise distribution to the target image distribution. Compared to GANs, they are more adapted to various image domains [34, 46] with the limited amount of training data [11, 32, 34]. Our method explores a way of ground-view synthesis conditioning on a combination of geo-location and context information using ControlNet [46] and LORA [11].

Cross-view geo-localization It focuses on estimating the location and orientation of ground-view images based on given satellite images. Early works regard it as the image retrieval problem that finds the most similar satellite image from a database to determine the rough location of query images. These works focus on designing powerful handcrafted features [2, 21] or learning-based features [1, 30, 39, 40] to bridge the cross-view domain gap. [12, 50] further identified the accurate pixel location on satellite images corresponding to the query images by employing the Siamese network to regress the location coordinates. Recent works [16, 36] utilized the geometry guidance to project the ground-view images to the top-down view domain, where the location and orientation of query images can be robustly regressed.

3 Method

We present a novel pipeline designed to predict ground-view panorama images using a set of satellite images, as depicted in Fig. 2. Our main goal is to perform geometrically accurate projection of satellite textures to the ground view, encompassing subtle details of building facades to enhance ground-view synthesis. The proposed pipeline consists of four stages: the top-down view stage, projection stage, ground-view stage, and texture-guided generation stage. The details are described below.

3.1 Top-down View and Projection Stage

We follow the 2D-3D-2D way to project the top-down satellite images to ground level. The satellite 3D geometry is first derived via well-established stereo matching methods [5, 17, 27, 43]. We then perform a 2D-3D texture projection (known as RPC projection [37]) to transform the top-down satellite textures to 3D space

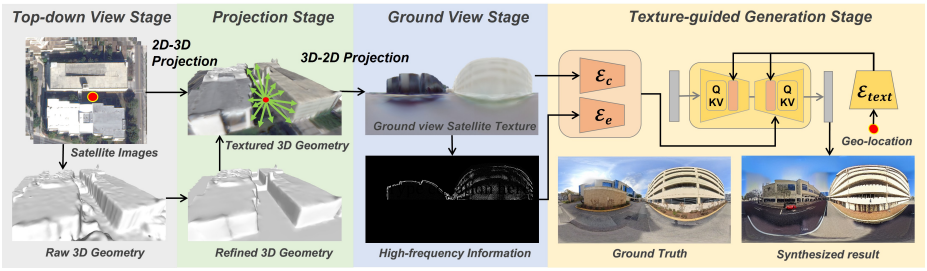


Fig. 2: Overview of our pipeline. **Top-down View Stage** and **Projection Stage**: the satellite textures are projected to the refined 3D geometry and then projected back to ground-view 2D space (Sec. 3.1). **Ground-view Stage:** The ground view satellite texture and corresponding high-frequency layout information serve as the conditions (Sec. 3.2). **Texture-guided Generation Stage:** We use the recent successful diffusion model [33] conditioning on ground-view satellite textures, high-frequency information with the geospecific prior. (Sec. 3.3)

.

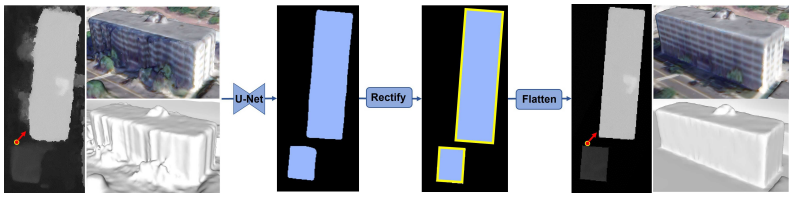
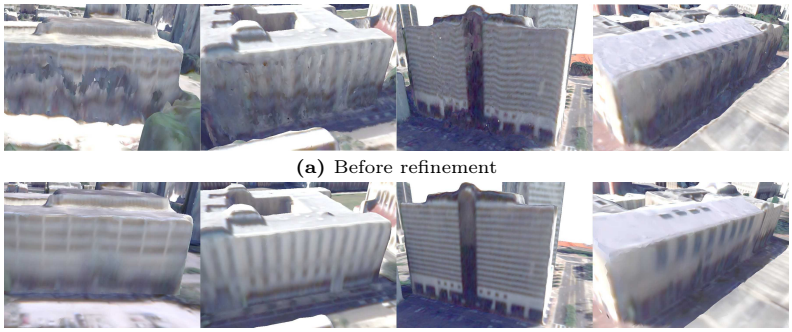


Fig. 3: Texture-friendly geometry refinement process. The process takes the original height map as input and estimates the building footprint, followed by boundary regularization to produce the refined height map.

and fuse the multiple overlap textures by optimizing the global illumination and color consistency. Then a panoramic projection is performed from the 3D texture information to the ground level. The crucial factor in achieving perfect 2D-3D projection lies in the precise and smooth geometry. Although the derived 3D geometry is mathematically computed based on multi-view constraints, the presence of satellite sensor noises can introduce distortions that will largely impact the quality of projected textures around building facades. Therefore, we propose an effective approach to refine the geometry of satellite buildings.

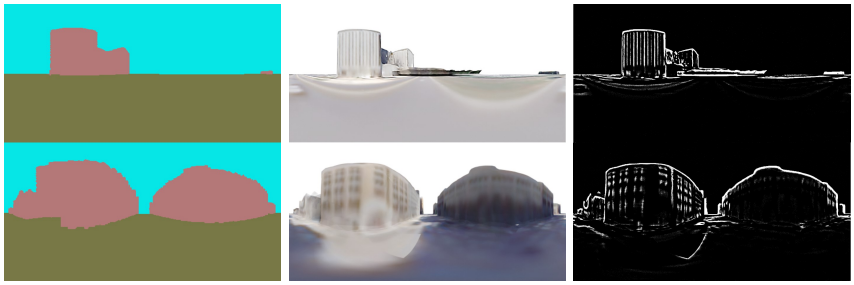
Texture-friendly geometry refinement. As the building facades are nearly perpendicular to the satellite viewing direction, even minor disturbances around the facade surface can lead to inaccuracies in mapping satellite textures to the correct facade location. To ensure the smoothness and precision of the satellite building geometry, we employ a 2D U-Net [42] to ascertain the high-level building footprint from satellite images, classifying each pixel as either part of a building or non-building region, as shown in Fig. 3. Subsequently, we detect and rectify building boundaries into a series of polygons, which will provide smooth building boundaries. For non-building regions, we conduct plane fitting and flat-



(a) Before refinement

(b) After refinement

Fig. 4: Examples of the satellite textures before and after our transformation-friendly geometry refinement.



(a) Semantics **(b)** Our satellite texture **(c)** Our edge map

Fig. 5: Illustration of three conditions for cross-view synthesis. Semantics are widely used by existing works [2, 22, 41]. Our satellite textures can provide additional high-frequency and color information that details the building facade layouts, such as the window/door shape and locations.

ten the non-building pixels onto the fitted plane, while retaining the original configuration of building pixels. Once the satellite geometry is refined, we perform 2D-3D texture projection, and then panoramic projection to derive the ground-view satellite textures. Our refinement method shows excellent results on various buildings, examples as shown in Fig. 4.

3.2 Ground-view Stage

After minimizing the cross-viewport difference between satellite and ground-view images, the remaining challenges primarily involve resolution and color, with the resolution exhibiting a difference of over 10 times. To address this resolution disparity, we introduce a novel texture-guided condition to enhance the informativeness of ground-view generation.

Texture-guided condition. Most existing cross-view synthesis works [2,22, 41] conditions on semantics, which are assumed as the known information [2],

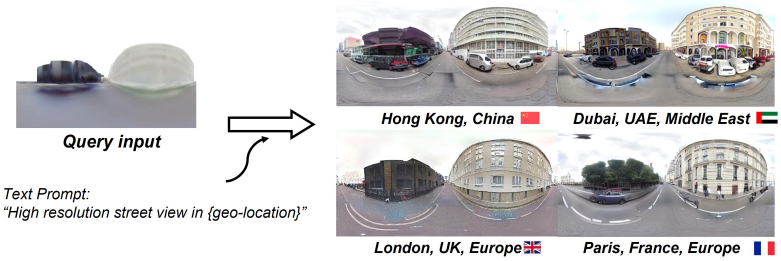


Fig. 6: Examples of the synthesized results of different geospecific priors conditioning on the same query condition.

internally estimated [41] or by 2D-3D-2D projection [22]. Deriving the semantics from satellite imagery is difficult and the semantic labels are limited to certain classes such as "Building", "Sky", "Road/Ground" and "Trees", limiting the diversity and richness of the details. Our semantics are generated based on top-down view building footprint segmentation results and perform 2D-3D-2D projection detailed in Sec. 3.1. In addition, we extract non-categorical, high-frequency information to preserve small granular structural details. Utilizing a 2D U-Net [3], we extract the building facade layout information, as shown in Fig. 5.

3.3 Texture-guided Generation Stage

We apply the latent diffusion model [33] as the base generator, which iteratively performs the denoising process from 2D random noise maps to synthesize the ground-view images.

Geospecific information prior. Many small-scale geographical nuances, such as specific vegetation types (e.g., palm trees in tropical regions), and diverse building facade features (including billboards and neon lights in Hong Kong), cannot be adequately captured by pure satellite texture. To address this, we embed the geospecific prior as the additional learnable parameters [11, 35] to the cross-attention module of our diffusion model. The geospecific information is represented as text descriptions corresponding to specific countries or regions, aligned with a set of street-view images from those areas to present the typical landscape of such regions. Specifically, given a geo-specific text prompt "High resolution street view in {geospecific}" $\mathbf{P}$, we obtain a conditioning vector $\mathbf{c}_p = \mathcal{E}_{text}(\mathbf{P})$. Subsequently, we incorporate geo-location tokens $\mathbf{P}_t$ (e.g. HongKong, Dubai, Paris) into the prompt $\mathbf{P}$ to produce geospecific conditioning vector $\mathbf{c}_t$, formalized as:

$$\mathbf{c}_t = \mathcal{E}_{text}(\mathbf{P} + \mathbf{P}_t) \tag{1}$$

For the pre-trained weights of our diffusion model $W_0 \in R^{d\times k}$, where k is the dimension for the input feature vectors and d is the dimension of the output feature vectors, we follow [11] to introduce two low-rank matrices $A \in \mathbb{R}^{d\times r}, B \in \mathbb{R}^{r\times k}$

and rank $r \ll min(d, k)$. During the training process, using random Gaussian initialization for A and zero initialization for B, and the optimizer only optimizes these two matrices, the output h pass pre-trained weight W_0 will be modified as:

$$h = W_0 x + ABx \tag{2}$$

where x is the feature vector in the diffusion model. Fig. 6 illustrates that our method with geo-specific priors can generate images with geospecific attributes.

Texture Encoding. The ground-view satellite texture and edge map are encoded by another two networks [46] sharing the same architecture as our diffusion model, which conditions on such features by zero-convoluting them with each layer of the base model decoder. Given an input pair of images $\{\mathbf{z}_0, \mathbf{c}\}$, where $\mathbf{z}_0$ is the real street-view image, $\mathbf{c}$ are ground-view satellite textures. We first convert the condition images into feature space following the VQ-GAN [6] pre-processing pipeline.

$$\mathbf{c}_s = \mathcal{E}(\mathbf{c}). \tag{3}$$

where $\mathbf{c}_s$ is the texture condition vector that represents the color information of the satellite image. To incorporate the building structure information, we extract the edge condition vector $\mathbf{c}_e$ from $\mathbf{c}$

$$\mathbf{c}_e = \mathcal{E}(\mathcal{H}(\mathbf{c})). \tag{4}$$

where $\mathcal{H}$ is the edge map extraction network described in Sec. 3.2.

During the training process, given a time step $\mathbf{t}$ and a geospecific prompt (encoded as feature vector $\mathbf{c}_t$, see Eq. (1)), our diffusion-based network progressively adds Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ to the previous image $\mathbf{z}_{t-1}$ and produces a new noisy image $\mathbf{z}_t$ and it learns to predict the noise by minimizing the mean-square error:

$$\mathcal{L} = \mathbb{E}_{z_0, t, \mathbf{c}_s, \mathbf{c}_e, \mathbf{c}_t \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, \mathbf{c}_s, \mathbf{c}_e, \mathbf{c}_t)\|_2^2 \right] \tag{5}$$

Where $\mathcal{L}$ is the learning objective applied in our proposed approach. We aim to finetune the two texture encoding networks, where the first is conditioned by ground-view satellite RGB images, and the second is conditioned by the edge map.

3.4 Implementation Details

Datasets. We perform our experiments on a large-scale dataset, DFC 2019 [20], consisting of multi-view satellite images covering a $177km^2$ area in Jacksonville, USA. Examples of satellite data and associated products are shown in Fig. 7, based on which we process into the 3D models, and have collected corresponding ground-views from Google street-view. Specifically, we first conducted stereo matching [9, 27] and our texture-friendly geometry refinement as described in Sec. 3.1 to produce refined satellite geometry in the form of a height map. Ground-view depth maps were then generated through 3D-2D projection from the refined satellite geometry. Subsequently, we performed 2D-3D-2D projection

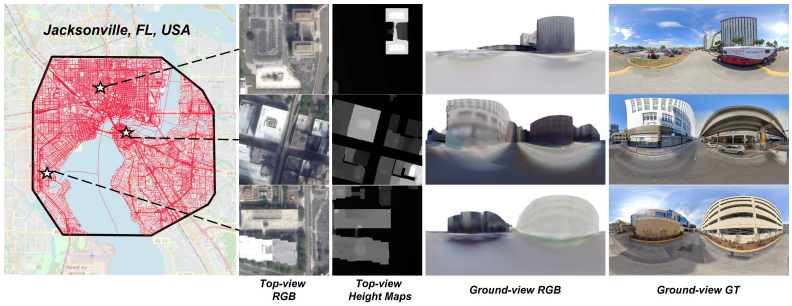


Fig. 7: Examples of our cross-view dataset based on Jacksonville, Florida, USA. It provides well-aligned pairs of ground-view satellite RGB and ground truth images, along with the top-view RGB and height maps.(see details in Sec. 3.4)

to map the top-down satellite texture to ground level. Semantic information was derived from OpenStreetMap Building Footprint data [24]. To obtain the reference ground view data, we collected Google street-view images within the study area using Google StreetView 360, with a step distance of 30 meters. Each image included location information (longitude, latitude, orientation). To address positional errors, we adopted a preprocessing strategy similar to [22], calculating the overlap ratio between sky regions in Google street-view images and ground-view satellite images. We selected pairs with an overlap ratio exceeding 95%, followed by post-processing. This process yielded over 7,000 pairs of cross-view data, each comprising top-down view RGB, height map, ground-view RGB, depth map, and ground truth images. For dataset split, we spatially tiled the datasets into a set of $700m \times 700m$ sub-tiles, with the ratio of train/val/test tiles being $8:1:1$. To prevent spatial correlation and overfitted prediction, we selected our test tile far apart from the training tiles.

Training Scheme. For the building footprint segmentation network, we randomly cropped out 512 x 512 patches from satellite images and assigned each pixel with 1 or 0 representing whether belongs to the building or not based on OpenStreetMap data with post-processing. This building footprint dataset is self-constructed based on the DFC-2019 multi-view images, which contain 11,784 pairs. Given its moderate size and simplicity of the task (predicting a binary task), We chose a moderately complex network, SegFormer [42] as our segmentation network, which was trained for 40 epochs.

For the texture-guide generation stage, we used pre-trained weights from Stable Diffusion v1.5 [33] as the base diffusion model and only finetuned the newly added parameters for each geospecific text prompt (detailed in Sec. 3.3). The paired data for geospecific prior training was a text prompt of "High-resolution street view in <City, Country, Continent>" and a set of Google street-view images broadly in the region to reflect the types of buildings and city landscape. In our experiment, we encoded five cities (but easily expandable) including London, Hong Kong, Jacksonville, Paris, and Dubai, where each city contains around 500

Table 1: Quantitative evaluation of synthesized image quality. We compared our method with Sat2Ground [22], Sat2Density [26], CrossMLP [31] and PanoGAN [41]. The same metrics for Tab. 2.

Method	**Low level**		**Edge level**	**Semantic level**			**Perceptual Level**		
	PSNR↑	SSIM↑	I_E↑	I_B↑	I_G↑	I_S↑	LPIPS↓	FID↓	DreamSIM↓
Sat2G	**21.02**	0.388	0.072	0.345	0.324	0.851	0.527	160.6	0.420
Sat2D	19.04	0.388	0.067	0.285	0.310	0.782	0.574	227.3	0.481
CrossMLP	18.66	**0.407**	0.069	0.448	0.214	0.861	0.509	170.8	0.434
PanoGAN	20.51	0.373	0.078	0.376	0.457	0.801	0.488	98.81	0.348
Ours	19.95	0.397	**0.089**	**0.570**	**0.864**	**0.874**	**0.449**	**71.04**	**0.315**

images. For ground-view texture encoders, we created the trainable copy of the UNet encoder according to ControlNet [46] and trained the encoders with the train datasets and inference on test datasets.

4 Experiments

4.1 Baselines and Metrics

Baseline methods. We chose two direct synthesis methods, **CrossMLP** [31] and **PanoGAN** [41] and two geometry-guided methods **Sat2Ground** [22] and **Sat2Density** [26] as baseline methods. The implementation and modification can be found in supplementary material.

Evaluation metrics. We use a combination of low-level, structure-level, semantic-level, and perceptual-level metrics to evaluate the quality of synthesized results. **Low-level metrics**. We follow [22, 29, 41] and use PSNR, and SSIM, which evaluate differences per pixels or local patches. **Edge-level metrics**. We extract the edge map from the synthesized result and ground truth using the Canny detector and calculate their average IoU (intersection over union) as the edge level similarity metric, denoted as I_e. **Semantic level metrics**: We calculate the average IoU of building (I_B), ground (I_G), and sky labels (I_S) between synthesized and ground truth images, where the semantic labels are generated from OneFormer [15] trained on ADE20K dataset [48]. **Perceptual level metrics**: we apply three widely-used perceptual metrics: the Fréchet Inception Distance (FID) [8] and the Learned Perceptual Image Patch Similarity (LPIPS) [47], and DreamSIM [7].

4.2 Comparison to State-of-the-Art Methods

Tab. 1 and Fig. 8 provide quantitative and qualitative comparison results of CrossMLP [31], PanoGAN [41], Sat2Ground [22], Sat2Density [26] and ours in the dataset described in Sec. 3.4.

(a) Sat2G **(b)** Sat2D **(c)** CrossMLP **(d)** PanoGAN **(e)** Ours **(f)** GT

Fig. 8: Qualitative comparison. We present various synthesis results of our method, compared with Sat2Ground [22], Sat2Density [26], CrossMLP [31], PanoGAN [41]. Our results are more photorealistic than the baseline methods.

Semantic level similarity. For a semantic category, a higher mIoU indicates the synthesized objects are more recognized by the pre-trained semantic segmentation models. For building objects, ours achieved 25.8% and 54.8% improvement than CrossMLP and PanoGAN. Similarly for road objects, ours achieved 685% and 402% improvement than CrossMLP and PanoGAN. Specifically, Sat2Ground synthesized the basic building layouts because of its geometry-guided module while the synthesized building objects were hardly recognized by the pre-trained model, which can be attributed to its GAN-based synthsis module. Moreover, the road regions generated by CrossMLP and PanoGAN exhibit blurry and repetitive artifacts, which are misclassified as 'Sand' and 'Earth' categories, and a similar issue occurs with the "Building" label.

Perceptual level similarity. As shown in Fig. 8, our approach produced fewer artifacts, and the synthesized building facades were more similar to those

Table 2: Quantitative ablation study of the proposed geospecific priors, the ground-view satellite texture condition, termed as "RGB". The same settings for Fig. 9.

Method	**Low level**		**Edge level**	**Semantic level**			**Perceptual Level**		
	PSNR↑	SSIM↑	I_E↑	I_B↑	I_G↑	I_S↑	LPIPS↓	FID↓	DreamSIM↓
Ours	**19.82**	**0.389**	**0.090**	**0.550**	**0.898**	**0.876**	**0.456**	**66.06**	**0.309**
w/o prior	17.34	0.288	0.071	0.359	0.832	0.726	0.580	87.75	0.416
w/o RGB	18.74	0.350	0.086	0.521	0.722	0.862	0.587	89.60	0.324

in the ground truth images. As shown in Tab. 1, ours achieved better FID results with more than 15.7% compared to baseline methods. The baseline methods generated artifacts of varying degrees around the building and ground areas. CrossMLP, in particular, synthesized buildings with transparent effects that blend with the sky and vegetation. PanoGAN and Sat2Ground performed slightly better than CrossMLP around building regions, displaying the basic facade layouts. However, these layouts appear to be more randomly generated and contain numerous repetitive artifacts.

Edge and low-level similarity. For edge-level performance, ours achieved an improvement of more than 15.5% than baseline methods. This superiority was attributed to the ground-view satellite texture conditions offering more high-frequency information compared to semantics. For low-level performance, ours ranked second in SSIM and third in PSNR. Given that these metrics evaluate differences at the per-pixel or patch level, our diffusion-based synthesis model is not designed to precisely replicate dataset distributions at the pixel level. Instead, it focuses on synthesizing structures and perceptual features. Additionally, these metrics are particularly sensitive to the inherent randomness of diffusion-based models, which arises from their training on large-scale datasets. For instance, in Fig. 8, rows 3, 6, 7, and 9 of our synthesized results exhibit significant deviations from the ground truth at the pixel level, including variations in lighting, shadows, clouds, and tree shapes.

4.3 Ablation Study

We further investigate the influence of multiple key components of our pipelines.

Importance of geo-lcoation prior. We removed the geospecific prior, formulated as the additional parameters in the cross-attention module of the diffusion model, and performed the training to see the influence of geospecific prior. As shown in Tab. 2, ours using geospecific prior achieved the improvement of 32.8% and 25.7% in FID and DreamSIM, 53.2% in $mIoU$ of the "Building" label than ours without prior.

Importance of texture condition.We substituted ground-view satellite textures with semantics and trained our models to assess the significance of texture conditions. As shown in Tab. 2, our approach utilizing texture conditions surpassed the one employing semantics by 23.5%, 20.7%, and 4.6% across

(a) Ours w/o RGB **(b)** Ours w/o prior **(c)** Ours **(d)** GT

Fig. 9: Qualitative ablation study. We show the visual comparison of the synthesized images by our methods with different configurations.

Table 3: Quantitative results for the potential improvement of baseline methods with our ground-view satellite texture conditions over semantics.

Method	LPIPS↓		FID↓		DreamSIM↓	
	Semantics	Textures	Semantics	Textures	Semantics	Textures
CrossMLP [31]	**0.509**	0.535	170.8	**86.65**	0.434	**0.432**
PanoGAN [41]	0.488	**0.477**	98.81	**84.38**	0.348	**0.337**
Ours	0.587	**0.449**	89.60	**71.04**	0.324	**0.315**

three perceptual-level metrics. The ground-view satellite texture provides richer details, thereby establishing a less ambiguous mapping relation. Moreover, the qualitative results depicted in Fig. 9 underscore the challenges faced by our semantics-dependent approach in accurately synthesizing building facade layouts and textures.

Potential improvement for baselines. To showcase the effectiveness of the proposed ground-view satellite textures, we conducted an experiment comparing them with two baseline methods, CrossMLP [31] and PanoGAN [41], which originally utilize ground-view semantics as auxiliary inputs. Instead of semantics, we integrated ground-view textures and assessed their perceptual performance, as illustrated in Fig. 9. Overall, with our ground-view texture integration, both baseline methods demonstrated improved performance. Specifically, PanoGAN exhibited enhancements of 2.2%, 14.6%, and 3.1% in LPIPS, FID, and DreamSIM metrics, respectively. CrossMLP, when augmented with texture conditions, showed superior performance in FID and DreamSIM but slightly inferior results in the LPIPS metric.

4.4 Limitation

Although the synthesized views are geospecific with the help of geospecific prior and ground-view satellite textures, they currently lack view consistency among

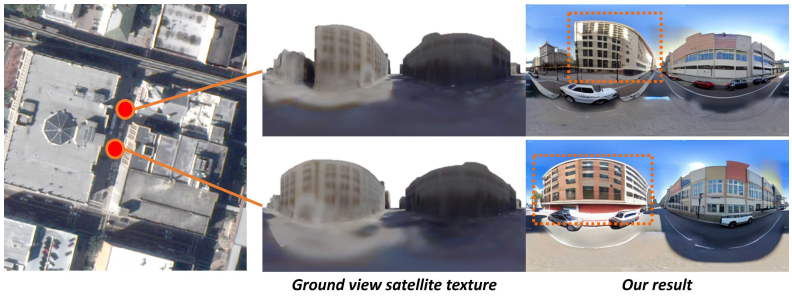


Fig. 10: Examples of our limitation. The inherent randomness of the diffusion model makes the synthesis results (marked as orange rectangles) not consistent with their neighbor views.

neighboring views, as shown in Fig. 10. For instance, buildings synthesized at two adjacent locations (highlighted by orange rectangles) show inconsistencies in color and facade layouts. This issue arises because the inherent randomness of the diffusion model prevents the results from strictly adhering to the input conditions in the absence of explicit cross-location or view-consistency constraints. Addressing this limitation to produce not only photorealistic but also consistent view sequences is a focus for future work.

5 Conclusion

In this paper, we propose a novel pipeline for predicting ground-view images from multi-view satellite images. The predicted ground views are geospecific, in that the generated views are not only consistent with the geometry derived from the satellite views but also the textural information from the satellite view, with a resolution enhancement at a factor of 10 or more. This stands for our work as the first that achieves view prediction that is specifically real to geolocation. In particular, we propose a geometry refinement module to refine satellite 3D geometry, to minimize the texture distortions on building facades, which significantly improves the transfered structural information from the weak satellite textures to the predicted views. Moreover, we propose to use a geospecific prior, to control the learned distributions of the diffusion model, to predict views that respect the local street-view styles. This encoded geospecific prior not only distills the generation to be geospecific but is also shown to be extremely effective in accelerating the training convergence. Our experiments demonstrate that our method significantly outperforms published baselines at a large margin, and is able to predict high-resolution, authentic ground views merely using multi-view satellite images.

Acknowledgements

This work is partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0034. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. It is also supported by the Office of Naval Research (Award No. N000142012141 and N000142312670). The authors would like to thank Xi Liu for his valuable discussion during this work.

References

1. Cai, S., Guo, Y., Khan, S., Hu, J., Wen, G.: Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8391–8400 (2019)
2. Castaldo, F., Zamir, A., Angst, R., Palmieri, F., Savarese, S.: Semantic cross-view matching. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 9–17 (2015)
3. Chan, C., Durand, F., Isola, P.: Learning to generate line drawings that convey geometry and semantics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7915–7925 (2022)
4. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
5. De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.M., Facciolo, G.: An automatic and modular stereo pipeline for pushbroom images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **2**, 49–56 (2014)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12873–12883 (June 2021)
7. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
9. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence **30**(2), 328–341 (2007)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

12. Hu, S., Feng, M., Nguyen, R.M.H., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
13. Huang, D., Tang, Y., Qin, R.: An evaluation of planetscope images for 3d reconstruction and change detection–experimental validations with case studies. GIScience & Remote Sensing **59**(1), 744–761 (2022)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
15. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998 (2023)
16. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.: Slicematch: Geometry-guided aggregation for cross-view pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17225–17234 (2023)
17. Leotta, M.J., Long, C., Jacquet, B., Zins, M., Lipsa, D., Shan, J., Xu, B., Li, Z., Zhang, X., Chang, S.F., Purri, M., Xue, J., Dana, K.: Urban semantic 3d reconstruction from multiview satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
18. Li, Z., Li, Z., Cui, Z., Pollefeys, M., Oswald, M.R.: Sat2scene: 3d urban scene generation from satellite images with diffusion. arXiv preprint arXiv:2401.10786 (2024)
19. Li, Z., Li, Z., Cui, Z., Qin, R., Pollefeys, M., Oswald, M.R.: Sat2vid: street-view panoramic video synthesis from a single satellite image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12436–12445 (2021)
20. Lian, Y., Feng, T., Zhou, J., Jia, M., Li, A., Wu, Z., Jiao, L., Brown, M., Hager, G., Yokoya, N., Hänsch, R., Saux, B.L.: Large-scale semantic 3-d reconstruction: Outcome of the 2019 ieee grss data fusion contest—part b. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 1158–1170 (2021). https://doi.org/10.1109/JSTARS.2020.3035274
21. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2013)
22. Lu, X., Li, Z., Cui, Z., Oswald, M.R., Pollefeys, M., Qin, R.: 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)
23. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
24. OpenStreetMap contributors: Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org (2017)
25. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
26. Qian, M., Xiong, J., Xia, G.S., Xue, N.: Sat2density: Faithful density learning from satellite-ground image pairs. arXiv preprint arXiv:2303.14672 (2023)

27. Qin, R.: Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **3**, 77–82 (2016)
28. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Advances in neural information processing systems **29** (2016)
29. Regmi, K., Borji, A.: Cross-view image synthesis using conditional gans. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3501–3510 (2018)
30. Regmi, K., Shah, M.: Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 470–479 (2019)
31. Ren, B., Tang, H., Sebe, N.: Cascaded cross mlp-mixer gans for cross-view image translation. arXiv preprint arXiv:2110.10183 (2021)
32. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on Graphics (TOG) **42**(1), 1–13 (2022)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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)
36. Shi, Y., Wu, F., Perincherry, A., Vora, A., Li, H.: 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. pp. 21516–21526 (2023)
37. Singh, S.K., Naidu, S.D., Srinivasan, T., Krishna, B.G., Srivastava, P.: Rational polynomial modelling for cartosat-1 data. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **37**(Part B1), 885–888 (2008)
38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
39. Vo, N.N., Hays, J.: 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. pp. 494–509. Springer (2016)
40. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
41. Wu, S., Tang, H., Jing, X.Y., Zhao, H., Qian, J., Sebe, N., Yan, Y.: Cross-view panorama image synthesis. IEEE Transactions on Multimedia (2022)
42. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)

43. Xu, N., Qin, R.: Large-scale dsm registration via motion averaging. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **10**, 275–282 (2024)
44. Xu, N., Qin, R., Huang, D., Remondino, F.: Multi-tiling neural radiance field (nerf)—geometric assessment on large-scale aerial datasets. The Photogrammetric Record (2024)
45. Xu, N., Qin, R., Song, S.: Point cloud registration for lidar and photogrammetric data: A critical synthesis and performance analysis on classic and deep learning algorithms. ISPRS open journal of photogrammetry and remote sensing **8**, 100032 (2023)
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
48. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
49. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
50. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)