Sat2Density: Faithful Density Learning from Satellite-Ground Image Pairs

Ming Qian[1,2] Jincheng Xiong[1] Gui-Song Xia[1] Nan Xue[2*]
[1]LIESMARS, Wuhan University [2]Ant Group
https://sat2density.github.io

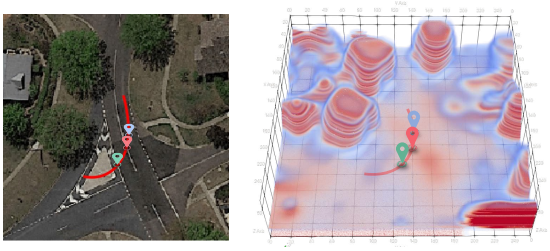
Abstract

This paper aims to develop an accurate 3D geometry representation of satellite images using satellite-ground image pairs. Our focus is on the challenging problem of 3D-aware ground-views synthesis from a satellite image. We draw inspiration from the density field representation used in volumetric neural rendering and propose a new approach, called Sat2Density. Our method utilizes the properties of ground-view panoramas for the sky and non-sky regions to learn faithful density fields of 3D scenes in a geometric perspective. Unlike other methods that require extra depth information during training, our Sat2Density can automatically learn accurate and faithful 3D geometry via density representation without depth supervision. This advancement significantly improves the ground-view panorama synthesis task. Additionally, our study provides a new geometric perspective to understand the relationship between satellite and ground-view images in 3D space.

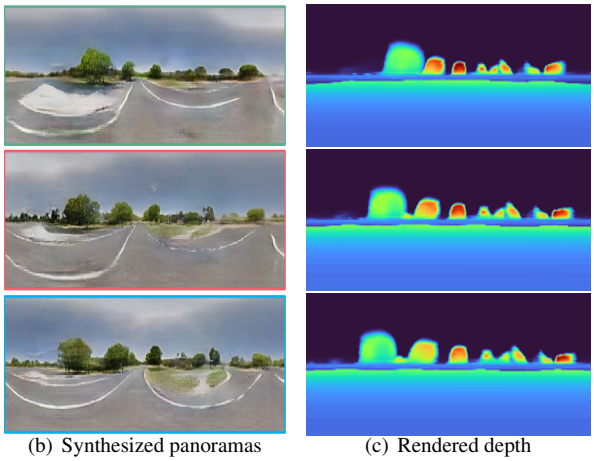
1. Introduction

The emergence of satellite imagery has significantly enhanced our daily lives by providing easy access to a comprehensive view of the planet. This bird's-eye view offers valuable information that compensates for the limited perspective of ground-level observations by humans. However, what specific information does satellite imagery provide, and why is it so crucial? In this paper, we propose that the most critical insights come from the analysis of the geometry, topology, and geography of cross-view observations captured by paired satellite and ground-level images. Building on this hypothesis, we aim to address the challenging problem of synthesizing ground-level images from paired satellite and ground-level imagery by leveraging density representations of 3D scenes.

The challenge of generating ground-level images from satellite imagery is tackled by leveraging massive datasets containing both satellite images and corresponding ground-level panoramas captured at the same geographical coordinates. However, the drastic differences in viewpoint between the two types of images, combined with the limited overlap of visual features and large appearance variations, create a highly complex and ill-posed learning problem. To address this challenge, researchers have extensively studied the use of conditional generative adversarial networks, which leverage high-level semantics and contextual information in a generative way [19, 20, 33, 26, 12]. However, since the contextual information used is typically at the im-



(a) Learned density from satellite images



(b) Synthesized panoramas (c) Rendered depth

Figure 1. Sat2Density trains with a collection of satellite-ground image pairs, without depth, or multi-view supervision. Our 3d GAN enables the synthesis of scenes conditioned on a satellite image, producing multi-view-consistent ground-view renderings and geometry. *Please see the project page for more videos.*

*Corresponding author

age level, the 3D information can only be marginally inferred during training, often resulting in unsatisfactory synthesis results.

Recent studies [22, 11] have suggested that accurate 3D scene geometry plays a crucial role in generating high-quality ground-view images. With extra depth supervision, Sat2Video [11] introduced a method to synthesize spatial-temporal ground-view video frames along a camera trajectory, rather than a single panorama from the center viewpoint of the satellite image. Additionally, Shi *et al*. [22] demonstrated that coarse satellite depth maps can be learned from paired data through multi-plane image representation using a novel projection model between the satellite and ground viewpoints, but a coarse 3d representation can not facilitate rendering 3D-aware ground-view images. Building on these insights, we aim to investigate whether it is possible to achieve even more accurate 3D geometry using the vast collection of satellite-ground image pairs.

Our study is motivated by the latest developments in the neural radiance field (NeRF) [16], which has shown promising results in novel view synthesis. Benefiting from the flexibility of density field in volumetric rendering [8], faithful 3D geometry can be learned from a large number of posed images. Therefore, we adopt density fields as the representation and focus on learning accurate density fields from paired satellite-ground image pairs. More precisely, in this paper, we present a novel approach called Sat2Density, which involves two convolutional encoder-decoder networks: DensityNet and RenderNet. The DensityNet receives satellite images as input to represent the density field in an explicit grid, which plays a crucial role in producing ground-view panorama images using the RenderNet. With such a straightforward network design, we delve into the goal of learning faithful density field first and then render high-fidelity ground-view panoramas.

While we employed a flexible approach to representing geometry using explicit volume density and volumetric rendering, an end-to-end learning approach alone is inadequate for restoring geometry using only satellite-ground image pairs. Upon examining the tasks and satellite-ground image pairs, we identified two main factors that may impede geometry learning, which has been overlooked in previous works on satellite-to-ground view synthesis. Firstly, the sky is an essential part of ground scenes but is absent in the satellite view, and it is nearly impossible to learn a faithful representation of the infinite sky region in each image using explicit volume density. Secondly, differences in illumination among the ground images during training make it challenging to learn geometry effectively.

With the above intuitive observation, we propose two supervision signals, the *non-sky opacity supervision* and *illumination injection*, to learn the density fields in a volumetric rendering form jointly. The *non-sky opacity supervision* compels the density field to focus on the satellite scene and ignore the infinity regions, whereas the *illumination injection* learns the illumination from sky regions to further regularize the learning density field. By learning the density field, our Sat2Density approach goes beyond the center ground-view panorama synthesis from the training data and achieves the ground-view panorama video synthesis with the best spatial-temporal consistency. As shown in Figure 1, our Sat2Density continuously synthesizes the panorama images along the camera trajectory. We evaluated the effectiveness of our proposed approach on two large-scale benchmarks [22, 34] and obtained state-of-the-art performance. Comprehensive ablation studies further justified our design choices.

The main contributions of our paper are:

- We present a geometric approach, Sat2Density, for ground-view panorama synthesis from satellite images in end-to-end learning. By explicitly modeling the challenging cross-view synthesis task in the density field for the 3D scene geometry, our Sat2Density is able to synthesize high-fidelity panoramas on camera trajectories for video synthesis without using any extra 3D information out of the training data.
- We tackle the challenging problem of learning high-quality 3D geometry under extremely large viewpoint changes. By analyzing the unique challenges that arise with this problem, we present two intuitive approaches *non-sky opacity supervision* and *illumination injection* to compel the density learning to focus on the relevant features in the satellite scene presented in the paired data while mitigating the effects of infinite regions and illumination changes.
- To the best of our knowledge, we are the first to successfully learn a faithful geometry representation from satellite-ground image pairs. We believe that not only do our new findings improve the performance of ground-view panorama synthesis, but the learned faithful density will also provide a renewed understanding of the relationship between satellite and ground-view image data from a 3D geometric perspective.

2. Related Work

2.1. Satellite-Ground Cross-View Perception

Both ground-level and satellite images provide unique perspectives of the world, and their combination provided us with a more comprehensive way to understand and perceive the world from satellite-ground visual data. However, due to the drastic viewpoint changes between the satellite and ground images, poses several challenges in geo-localization [34, 23, 24, 25], cross-view synthesis [22, 11,

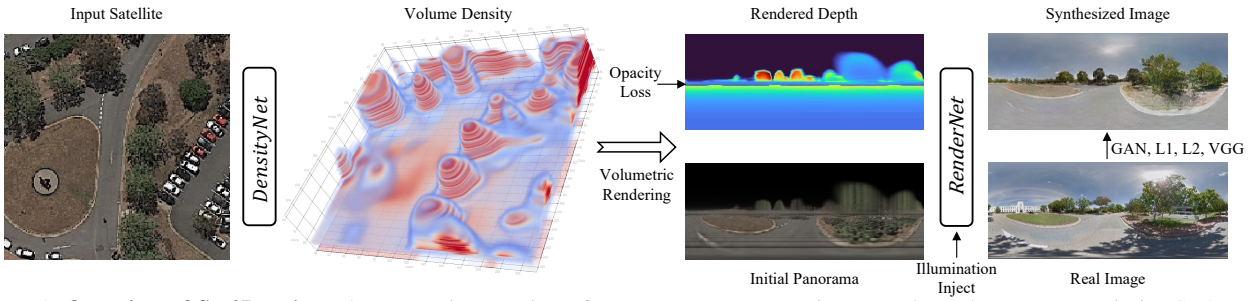


Figure 2. **Overview of Sat2Density.** The generation consists of two components, DensityNet and RenderNet. We optimize Sat2Density by reconstruction loss, adversarial loss, illumination injection loss, and opacity loss. *See text for details.*

12, 20, 26], overhead image segmentation with the assistance of ground-level images [30], geo-enabled depth estimation [29], predicting ground-level scene layout from aerial imagery [33].

To address this challenge, many previous works have proposed various approaches to model and learn the drastic viewpoint changes, including the use of homography transforms [20], additional depth or semantic supervision [26, 12, 11], transformation matrices [34], and geospatial attention [30], among others. Despite effectiveness, these approaches mainly address the challenge on the image level instead of the 3D scenes.

Most recently, Shi *et al*. [22] proposed a method to learn geometry in satellite scenes implicitly using the height (or depth) probability distribution map, which achieved better results in synthesized road and grassland regions through their geometry projection approach. However, their learned geometry has limited effectiveness as the rendered satellite depth cannot accurately recognize objects. We go further along the line to focus on the 3D scene geometry conveyed in the satellite-ground image pairs. We demonstrate that the faithful 3D scene geometry can be explicitly decoded and leveraged with an appropriate representation and supervision signals, to obtain high-fidelity ground-view panoramas. Besides, we believe that our study brings a novel perspective to rethink satellite-ground image data for many other challenging problems.

2.2. Neural Radiance Field

Benefiting from the flexibility of density field in volumetric rendering [16], faithful 3D geometry can be learned from a dense number of posed images [4, 13, 1, 3]. Recent works [2, 31, 32] based on NeRF have shown that 3D representation can be learned even with only a few views. In a co-current work [28], it is also pointed out that the flexibility of the density field helps to learn the 3D geometric structure from a single image by disentangling the color and geometry, which allows neural networks to capture reliable 3D geometry in occluded areas.

Our goal can be viewed as an extremely challenging problem of density-based few-view synthesis with extremely large viewpoint changes, which was not studied well in previous works. In our study, we demonstrated the possibility of learning faithful geometry in the volumetric rendering formulation, shedding light on the most challenging cross-view configurations for novel view synthesis.

3. The Proposed Sat2Density

Figure 2 illustrates the computation pipeline for our proposed Sat2Density. Given the input satellite image $I_{\text{sat}} \in \mathbb{R}^{H\times W\times 3}$ for the encoder-decoder DensityNet, we learn an explicit volume of the density field $V_\sigma \in \mathbb{R}^{H\times W\times N}$. We render the panorama depth and project the color of the satellite image along rays in the ground view to generate an initial panorama image and feed them to the RenderNet. To ensure consistent illumination of the synthesis, the histogram of color in the sky regions of the panorama is used as a conditional input for our method.

3.1. Density Field Representation

We encode an explicit volume density as a discrete representation of scene geometry and parameterize it using a plain encoder-decoder architecture in DensityNet G_{dns} to learn the density field:

$$V_\sigma = G_{\text{dns}}(I_{\text{sat}}) \quad \text{s.t. } V_{\cdot,\cdot,\cdot} \in [0,\infty). \tag{1}$$

where the density information $v = V(x,y,z)$ is stored in the volume of V for the spatial location (x,y,z). For any queried location that does not locate in the sample position of the explicit grid, tri-linear interpolation is used to obtain its density value. Suppose the size of the real-world cube is (X,Y,Z) in the satellite image, two corner cases are considered: 1) for the locations outside the cube, we set their density to zero, and 2) we set the density in the lowest volume (*i.e.*, V(x,y,z=0)) to a relatively large value (10^3 in our experiments), which made an assumption that all ground regions are solid.

With the density field representation, the volumetric rendering techniques [14] are applied to render the depth $\hat{d}$ and

opacity $\hat{O}$ along the queried rays by

$$\hat{d} = \sum_{i=1}^{S} T_i \alpha_i d_i, \qquad \hat{O} = \sum_{i=1}^{S} T_i \alpha_i, \tag{2}$$

where d_i is the distance between the camera location and the sampled position, T_i denotes the accumulated transmittance along the ray at t_i, and α_i is the alpha value for the alpha compositing, written by

$$\alpha_i = 1 - \exp\left(-\sigma(\mathbf{x}_i)\delta_i\right) \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)\,. \tag{3}$$

Unlike NeRF [16] that learns the radiance field to render the colored images, we take a copy-paste strategy to compute the colored images by copying the color from the satellite image along the ray via bilinear interpolation for image rendering in

$$\hat{c}_{\text{map}} = \sum_{i}^{S} T_i \alpha_i c_i, \tag{4}$$

where $c_i = c(x_i, y_i, z_i) = I_{\text{sat}}(\frac{x_i}{S_x} + \frac{H}{2}, \frac{y_i}{S_y} + \frac{W}{2})$. S_x and S_y are the scaling factors between the pixel coordinate of the satellite image and the grid coordinate in V_σ. To keep the simplicity of our Sat2Density, we did not use the hierarchical sampling along rays for the computation of depth, colors, and opacity.

Thanks to the flexibility of volumetric rendering, for the end task of ground-view panorama synthesis, it is straightforward to render the ground-view depth, opacity, and the (initial) colored image. For the subsequent RenderNet, it takes the concatenated tensor of the rendered panorama depth, opacity, and colors as input to synthesize the high-fidelity ground-view images.

Learning density could draw precise geometry information of the scene, but it is hard to acquire real density information of the satellite scene only from the satellite-ground image pairs. In our work, we propose two supervisions: *non-sky opacity supervision* and *illumination injection* to improve the quality of the 3D geometry representation.

3.2. Supervisions from Sky/Non-Sky Separation

Non-Sky Opacity Supervision. We draw inspiration from the study of panorama image segmentation [15, 35], which treats the sky region as a meaningful category in the segmentation task. By taking the off-the-shelf sky segmentation model [35] to obtain the sky masks for the training panorama images, we tackle the *infinity issue* with a novel *non-sky opacity supervision* proposed. Based on our discussion in Sec. 1, the pseudo sky masks provide a strong inductive basis to faithfully learn the density fields for our proposed Sat2Density in a simple way.

Denoted by $\mathcal{R}$ and $\mathcal{R}'$ the non-sky/sky regions of the ground-view panorama, the loss function $\mathcal{L}_{\text{snop}}$ of our proposed non-sky opacity supervision reads to

$$\mathcal{L}_{\text{snop}} = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \hat{O}(\mathbf{r}) - 1 \right\|_1 + \sum_{\mathbf{r}\in\mathcal{R}'} \left\| \hat{O}(\mathbf{r}) \right\|_1 . \tag{5}$$

Illumination Injection from Sky Regions. While the density field works well on images of static subjects captured under controlled settings, it is incapable of modeling many ubiquitous, real-world phenomena for ground-view panorama synthesis. More importantly, due to the lacking of correspondence from the sky regions in ground-view images to the paired satellite image, we find that the variable illumination in the ground images is a key factor preventing the model to learn faithful 3D geometry.

Accordingly, we present an illumination injection from sky regions of the panorama. For the sake of simplicity of design, we choose the RGB histogram information in the sky regions as the illumination hints. In our implementation, we first cut out the sky part from the ground image, then calculate the RGB sky histogram with N bins. To further exploit the representational ability of sky histograms, we follow the style encoding scheme proposed in SPADE [17] to transform the sky histogram into a fixed-dim embedding, which allows our Sat2Density learn reliable information of complicated illumination from the sky histogram for the RenderNet. From our experiments, we also find that the injection of sky histogram into the RenderNet could further improve the quality of explicit volume density by encouraging the DensityNet to focus on the stuff regions rather than being disturbed by the per-image illumination variations issue.

By combining the above two approaches, we solve the per-image illumination variations and infinity issue, and let our model focus on learning the scene geometry relationship between satellite and ground view (see Figure 3). Besides, thanks to the proposed sky histogram illumination injection approach, our model could control illuminations.

3.3. Loss Functions

Sat2Density is trained with both reconstruction loss and adversarial loss. For reconstruction loss, we follow GAN-based syntheses works, using a combination of the perceptual loss [7], L1, and L2 loss. In the adversarial loss, we use the non-saturated loss [9] as the training objective. Besides, $\mathcal{L}_{snop}$ is used for opacity supervision. For illumination learning, we follow the SPADE [18] use a KL Divergence loss. Last, in the discriminator, we use the feature matching loss & a modified multi-scale discriminator architecture in [27]. Details can be found in the supplemental material.

4. Experiments

4.1. Implementation Details

We train our model with 256×256 input satellite images and output a $256 \times 256 \times 65$ explicit volume density and finally predict a 128×512 360° panorama image, which is the same as the setting in [22] for a fair comparison. The maximum height modeled by our implicit volume density is 8 meters, which is an empirical setup. We approximate the height of the street-view camera as 2 meters with respect to the ground plane, which follows Shi *et al*. [22]. The bins of histogram in each channel are 90. The model is trained in an end-to-end manner with a batch size of 16. The optimizer we used is Adam with a learning rate of 0.00005, and $\beta_1 = 0$, $\beta_2 = 0.999$. Using a 32GB Tesla V100, the training time is about 30 hours for 30 epochs. As for the architectures of DensityNet and RenderNet, they share most similarities with the networks used in Pix2Pix[6]. More details about the model architecture and training details can be found in the supplemental material.

4.2. Evaluation Metrics

We use several evaluation metrics to quantitatively assess our results. The low-level similarity measures include root-mean-square error (RMSE), structure similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and sharpness difference (SD), which evaluate the pixel-wise similarity between two images. We also use high-level perceptual similarity [36] for evaluation as in previous works. Perceptual similarity evaluates the feature similarity of generated and real images. We employ the pretrained AlexNet [10] and Squeeze [5] networks as backbones for evaluation, denoted as P_{alex} and P_{squeeze}, respectively.

4.3. Dataset for Ground View Synthesis

We choose CVUSA [34] and CVACT(Aligned) [22] datasets for comparison in the central ground-view synthesis setting, following Shi *et al*. [22]. CVACT(Aligned) is a well-posed dataset aligned in Shi *et al*. [22], with 360° horizontal and 180° vertical visualization in panorama ground truth. Hence, we selected it for controllable illumination visualization and controllable video generation. For the dataset CVUSA, we only use it for center-ground view synthesis as their upper and lower parts of the panoramic image are trimmed for the geo-localization task, and the number of trimmed pixels is unknown [34]. During training and testing, we considered the street-view panoramas in the CVUSA dataset as having a 90° vertical field of view (FoV), with the central horizontal line representing the horizon. CVACT(Aligned) contains 26,519 training pairs and 6,288 testing pairs, while CVUSA contains 35,532 training pairs and 8,884 testing pairs. We did not choose other available datasets built in the city scene, such as OP[21] and VIGOR[37], since their GPS information is less accurate in urban areas compared to open rural areas, and poorly posed satellite-ground image pairs are not suitable for our task.

4.4. Ablation Study

In this section, we conduct experiments to validate the importance of each component in our framework, including *non-sky opacity supervision*, *illumination injection*, and whether to concatenate depth with the initial panorama before sending it to the RenderNet. We first present quantitative comparisons in Table 1 for the center ground-view synthesis setting. It is evident that the illumination injection most affects the quantitative result, at the same time, only adding non-sky opacity supervision will lead to a little drop in the quantitative score. But combining the two approaches will lead to better scores. Moreover, the comparison on whether concatenate depth to the RenderNet shows almost equal results in terms of quantitative comparison.

Figure 3 shows some samples from the rendered panorama video and their corresponding depth maps. The results show that without the proposed components (baseline), the rendered depth seems meaningless in the upper half, while the lower regions look good. We attribute this phenomenon to the fact that the lower bound of the panorama is the ray that looks down, which is highly related to the ground region near the shooting center in the satellite. It can be easily learned by a simple CNN with a simple geometry projection, which also explains why the work in [22] can render the ground region well.

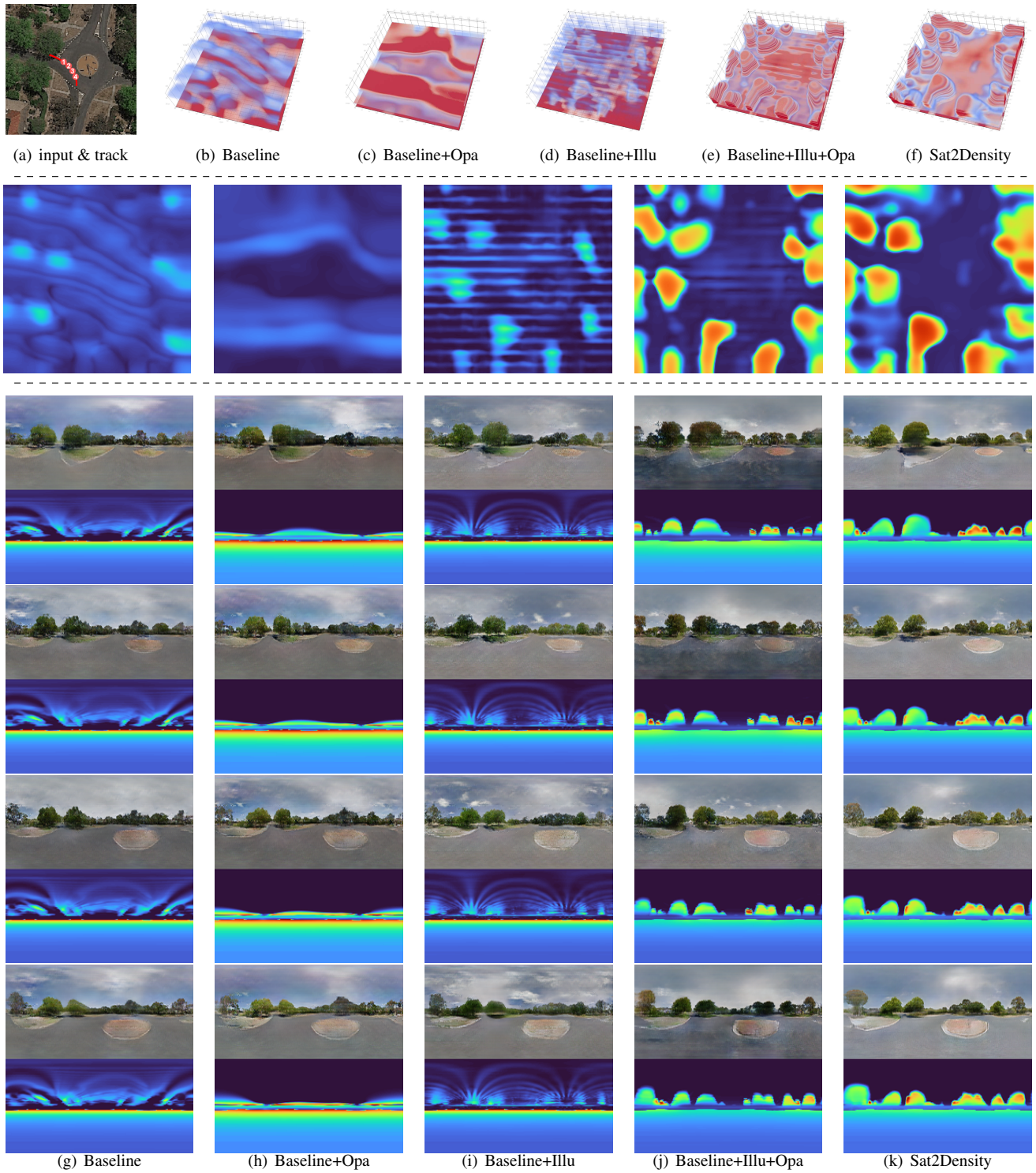
Compared to the baseline, adding the illumination injection can make the rendered depth look better, but the trees' density looks indistinct, and the sky region's density is still unclear. While only adding non-sky opacity supervision, the air region's opacity turns to zero, but the area between air and the ground is still barely satisfactory. The supervision did clear the sky region in the volume density, but the inner region between the sky and the ground is also smoothed. This is because such coarse supervision cannot help the model recognize the complex region.

By combining both strategies (Baseline+Illu+Opa), we can achieve a plausible 3D geometry representation that can generate a depth map faithful to the satellite image and the reference illumination. The volume density is clear compared to the above settings, and we can easily distinguish the inner regions.

Furthermore, when depth is incorporated into the rendering process, the resulting images tend to emphasize regions with depth information. This reduces the likelihood of generating objects in infinite areas randomly and leads to a synthesized ground view that more closely resembles the satellite scene, which can be observed from the video.

4.5. Center Ground-View Synthesis Comparison

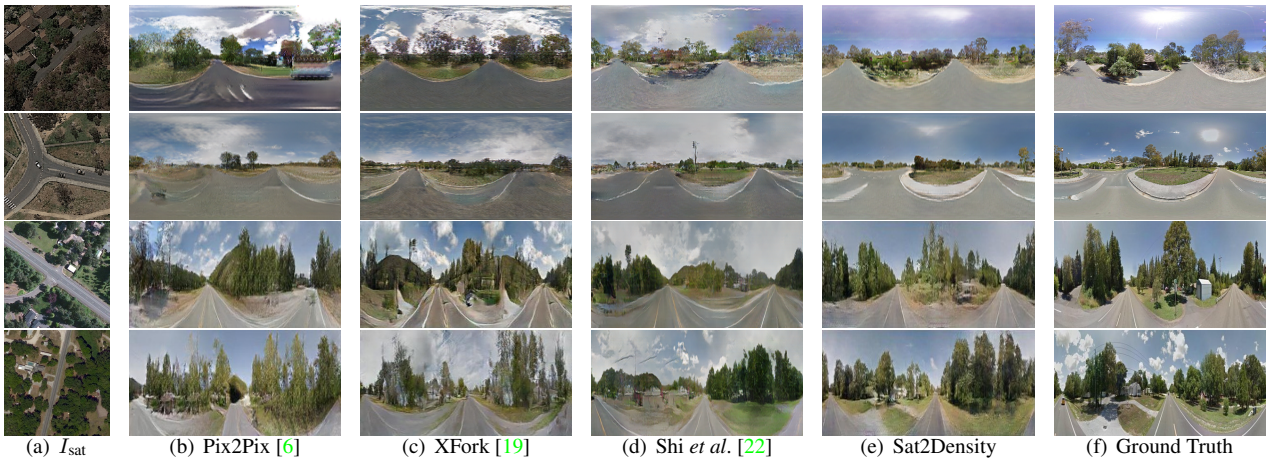
In the center ground-view synthesis setting, we compare our method with Pix2Pix [6], XFork [19], and Shi *et*



(a) input & track (b) Baseline (c) Baseline+Opa (d) Baseline+Illu (e) Baseline+Illu+Opa (f) Sat2Density

(g) Baseline (h) Baseline+Opa (i) Baseline+Illu (j) Baseline+Illu+Opa (k) Sat2Density

Figure 3. Ablation study on CVACT (Aligned) dataset. In the first row, the picture on the upper left is the input image. Each point from left to right is related to the bottom four rows from up to down. The remaining five images in the first row are the density rendered from the input satellite image following the setting (b-f) one by one. The images in the second row are the satellite depth calculated following the setting (b-f) one by one. 'Baseline' means baseline, 'Opa' means add *non-sky opacity supervision*, 'Illu' means add *illumination injection*, and 'Sat2Density' is our final result, compared to 'Baseline+Illu+Opa', we concatenate the depth map and initial panorama together to send to the RenderNet rather than only the initial panorama. *The video could be seen in the project page.*



(a) I_{sat} (b) Pix2Pix [6] (c) XFork [19] (d) Shi *et al*. [22] (e) Sat2Density (f) Ground Truth

Figure 4. Example images generated by different methods in center panorama synthesis task. The top two rows show the results on CVACT (Aligned) dataset, and the bottom two rows show the results on the CVUSA dataset.

Comparison	RMSE ↓	SSIM↑	PSNR↑	SD ↑	P_{alex} ↓	P_{squ} ↓
Base	48.40	0.4491	14.67	12.76	0.3772	0.2486
Base+Opa	48.39	0.4431	14.63	12.70	0.3847	0.2525
Base+Illu	41.62	0.4689	15.96	12.90	0.3497	0.2225
Base+Opa+Illu	40.71	0.4710	16.16	12.83	**0.3329**	0.2154
Sat2Density	**39.76**	**0.4818**	**16.38**	**12.90**	0.3339	**0.2145**

Table 1. Ablation study results on CVACT (Aligned) dataset. 'Base' means baseline, 'Opa' means add *non-sky opacity supervision*, 'Illu' means add *illumination injection*, and 'Sat2Density' is our result, compared to 'Baseline+Opa+Illu', we concatenate the depth map and initial panorama together to send to the RenderNet rather than only the initial panorama.

	Method	RMSE ↓	SSIM↑	PSNR↑	SD ↑	P_{alex} ↓	$P_{squeeze}$ ↓	Inference time/ms
CVACT (Ali.)	Pix2Pix [6]	49.75	0.3852	14.38	12.09	0.4654	0.3096	10.29
	XFork [19]	48.95	0.3710	14.50	12.32	0.4638	0.3262	17.24
	Shi *et al*. [22]	48.50	0.4272	14.59	12.31	0.4059	0.2708	109.88
	Sat2Density	47.13	0.4586	14.92	12.77	0.3842	0.2573	-
	Sat2Density-oracle	**39.76**	**0.4818**	**16.38**	**12.90**	**0.3339**	**0.2145**	33.12
CVUSA	Pix2Pix	55.27	0.2946	13.48	11.97	0.5092	0.3902	-
	XFork	54.11	0.2873	13.68	12.15	0.5144	0.4041	-
	Shi *et al*.	53.75	0.3451	13.75	12.06	0.4639	0.3506	-
	Sat2Density	53.30	0.3301	13.78	12.38	0.4504	0.3365	-
	Sat2Density-oracle	**48.75**	**0.3584**	**14.66**	**12.53**	**0.4163**	**0.3058**	-

Table 2. Quantitative comparison with existing algorithms on the CVACT (Aligned) and CVUSA datasets in center ground-view synthesis setting. 'Sat2Density' means we randomly choose nine histograms from the test set when inference and then average it. 'Sat2Density-oracle' means the histogram is from the GT ground image. The inference time was tested on a Tesla V100.

al. [22]. Pix2Pix and XFork are classic GAN-based models for image-to-image translation, but ignore the 3D geometry connections between the two views. Shi *et al*. [22] is the first geometry-guided synthesis model, which represents the 3D geometry in the depth probability MPI, showing brilliant results in the center ground-view synthesis setting.

Quantitative Comparison. As presented in Table 2, it is evident that Sat2Density achieves the best performance on all scores, including both low-level and perceptual similarity measures. Even choosing illumination hint randomly, our model still outperforms other methods.

Moreover, a combined analysis of the quantitative results of Sat2Density-sin and controllable illumination in Figure 5 reveals that illumination can significantly affect both common low-level and perceptual similarity measures, although the objects in the scene remain unchanged. As a result, it is more important to consider qualitative comparisons and video synthesis results.

Qualitative Comparison. In Figure 4, we find that the condition-GAN based methods can only synthesize good-looking ground images, but can not restore the geometry information from the satellite scene. Shi *et al*. [22] learn a coarse geometry representation, so the 3D information in the ground region is more reliable. For our method, as discussed in the ablation study, the high-fidelity synthesis (especially in the most challenging regions between the sky and the ground) is approached by learning faithful density representation of the 3D space.

4.6. Controllable Illumination

As shown in Figure 5, the sky histogram could easily control the image's illumination, while the semantics did not change, *e.g*. The road's color was changed by giving different illumination, but the shape remains unchanged. Besides, the illumination interpolation results can be seen on the project page, which also show the superiority of illumination injection.

4.7. Ground Video Generation

In Figure 6, we compare the rendered satellite depth, and synthesized ground images from a camera trajectory with the expansion of Shi *et al*. [22]. Shi *et al*. [22] focus on synthesizing ground panorama in the center of the satellite image, as they learn geometry by depth probability map, we expand their work by moving the camera position when

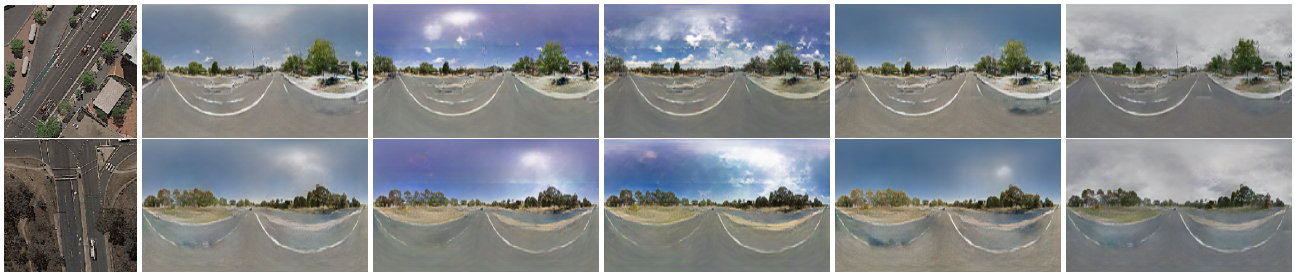
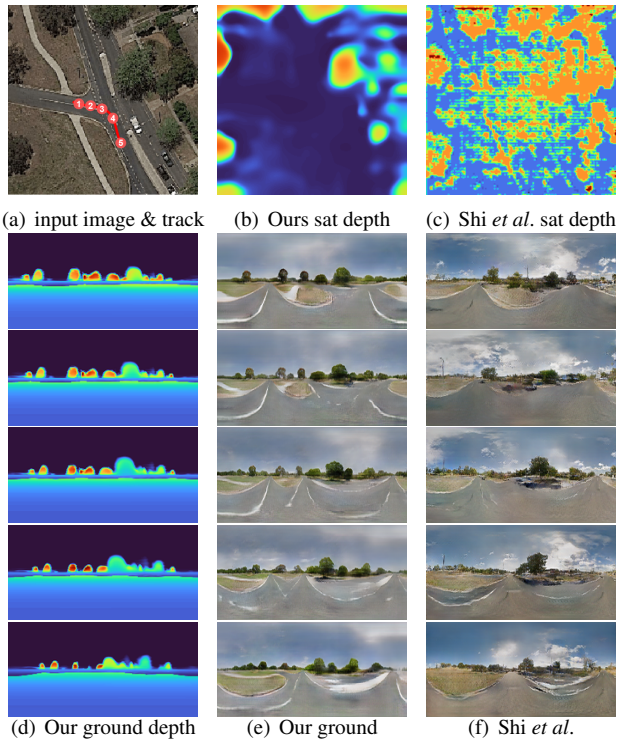


Figure 5. Controllable illumination synthesis: each row shows the results rendered from the same satellite image, and each column shares the same illumination from the same ground image.



(a) input image & track (b) Ours sat depth (c) Shi *et al*. sat depth

(d) Our ground depth (e) Our ground (f) Shi *et al*.

Figure 6. Synthesized video & depth comparison. (a) is the input satellite image, the red curve is the camera trajectory to synthesize video, the red point is chosen for visualization in (d-f), (b-c) is the rendered satellite depth by our method and Shi *et al*., (d) is the rendered depth of the ground image by our method, (e), (f) are the rendered ground image by our method and Shi *et al*. separately. *The video can be seen in the project page.*

inference. We also show the rendered depth maps. It is worth noting that Shi *et al*. [22] cannot generate a depth map for novel views, due to the intrinsic flaw of the depth probability representation.

From the synthesized satellite depth, we observe that Shi *et al*. [22] can only render a very coarse satellite depth, and is hard to recognize most regions. In contrast, trees and ground regions can easily distinguish from our satellite depth, and the depth in ground regions appears smooth. Additionally, we can render depth in any view direction by volume rendering, as shown in Figure 6 (d).

Furthermore, we find that the rendered ground video by the expanded Shi *et al*. [22] has little consistency due to the unfaithful 3D geometry representation, as evidenced by the inconsistencies present in the trees and sky. These results demonstrate that Sat2Density is capable of rendering temporal and spatially consistent videos.

5. Discussion and Limitations

Although Sat2Density is advancing the state-of-the-art, it still has some limitations. For instance, the tree density and visibility of houses are not perfect in our results, which would come down to the following reasons. Firstly, the one-to-one satellite and ground image pairs are not optimal for training, as having multiple ground images corresponding to one satellite image would be better. Additionally, images taken on different days may introduce transient objects that our approach is unable to handle. Then, the projected color map sent to the RenderNet may be too coarse in the region between sky and ground, which could impact the final result. Finally, well-aligned image pairs are required to learn geometry, so we are unable to evaluate the effectiveness of our approach in city scenes for more coarse GPS precision in the city. Therefore, while our work is a promising start for learning geometry from cross-view image pairs, there are still many challenges that need to be addressed.

6. Conclusion

In this paper, we propose a method, *i.e*. Sat2Density, to learn a faithful 3D geometry representation of satellite scenes from satellite-ground image pairs through the satellite-to-ground-view synthesis task. Our approach tackles two critical issues, the infinity issue and the illumination difference issue, to make geometry learning possible. By leveraging the learned density, our model is capable of synthesizing spatial and temporal ground videos from satellite images even with only one-to-one satellite-ground image pairs for training. To the best of our knowledge, our method represents the first successful attempt to learn precise 3D geometry from satellite-ground image pairs, which significantly advances the recognition of satellite-ground tasks from a geometric perspective.

Acknowledgements. This work was supported by the National Nature Science Foundation of China under grant 62101390.

References

[1] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv:2207.13751*, 2022. 3

[2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16102–16112, 2022. 3

[3] Terrance DeVries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 14284–14293, 2021. 3

[4] Zekun Hao, Arun Mallya, Serge J. Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 14052–14062, 2021. 3

[5] 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 $<$ 0.5mb model size. *arXiv:1602.07360*, 2016. 5

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5967–5976, 2017. 5, 7

[7] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9906, pages 694–711, 2016. 4

[8] James T. Kajiya and Brian Von Herzen. Ray tracing volume densities. In *ACM SIGGRAPH computer graphics*, pages 165–174, 1984. 2

[9] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 852–863, 2021. 4

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012. 5

[11] 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 IEEE International Conference on Computer Vision (ICCV)*, pages 12416–12425, 2021. 2, 3

[12] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 856–864, 2020. 1, 3

[13] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 7210–7219, 2021. 3

[14] Nelson L. Max. Optical models for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.*, 1(2):99–108, 1995. 3

[15] Radu Paul Mihail, Scott Workman, Zach Bessinger, and Nathan Jacobs. Sky segmentation in the wild: An empirical study. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–6, 2016. 4

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12346, pages 405–421, 2020. 2, 3, 4

[17] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2337–2346, 2019. 4

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2337–2346, 2019. 4

[19] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3501–3510, 2018. 1, 5, 7

[20] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Comput. Vis. Image Underst.*, 187, 2019. 1, 3

[21] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 470–479, 2019. 5

[22] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):10009–10022, 2022. 2, 3, 5, 7, 8

[23] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[24] 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. 2

[25] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11990–11997, 2020. 2

[26] 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 (CVPR)*, pages 2417–2426, 2019. 1, 3

[27] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8798–8807, 2018. 4

[28] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. *CoRR*, abs/2301.07668, 2023. 3

[29] Scott Workman and Hunter Blanton. Augmenting depth estimation with geospatial context. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 4542–4551, 2021. 3

[30] Scott Workman, Muhammad Usman Rafique, Hunter Blanton, and Nathan Jacobs. Revisiting near/remote sensing with geospatial attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 1768–1777, 2022. 3

[31] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13682, pages 736–753, 2022. 3

[32] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4578–4587, 2021. 3

[33] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4132–4140, 2017. 1, 3

[34] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4132–4140, 2017. 2, 3, 5

[35] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16896–16906, 2022. 4

[36] 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/CVF conference on computer vision and pattern recognition (CVPR)*, pages 586–595, 2018. 5

[37] Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3640–3649, 2021. 5