# Sat2Density: Faithful Density Learning from Satellite-Ground Image Pairs
## *Supplementary Material*

## Appendix

In the appendix, we first show our motivation and then present the details of the model architecture and training process of our Sat2Density model. After that, we describe the satellite and ground-view panorama camera models. Last, we give more discussion about our work for a better understanding of our paper and hope to advance viewing remote sensing and ground imagery from a geometric perspective. Code, pretrained models, and more video results can be found on our project page.

## A. Motivation

Sat2Density focuses on the geometric nature of generating high-quality ground street views conditioned on satellite images learning from collections of satellite-ground image pairs. The long-suffered issue from the unknown 3D information is addressed by separating the sky/non-sky regions with reasonable 3D density volumes learned. We believe our new perspective on the longstanding yet challenging problem of satellite-ground novel view synthesis would bring more insights for a wide range of 3D vision tasks, including but not limited to (1) using satellite images for autonomous driving with faithful 3D geometry, (2) providing promising and novel solutions for visual localization with satellite images.

## B. Addition Implementation Details

### B.1. DensityNet

The DensityNet is taken from the generator of Pix2Pix [2]. Compared to vanilla Pix2Pix in PyTorch implementations from pix2pix in PyTorch, our generator replaces the activation function in the initial layer and downsample layers from ReLU to PReLU, sets the number of resblock to 6, and replaces ReLU with Tanh in the last layer. The final output of DensityNet is an explicit volume density $V_\sigma \in \mathbb{R}^{H\times W\times N}$, rather than predicting an image with resolution $H \times W \times 3$.

### B.2. Illumination Injection

To inject the illumination, we first calculate the RGB histogram of the sky region in ground image with 90 bins in each color channel. Following the way process style in GANcraft[1], we use a style encoder to predict a style code, then use an MLP that is shared across all the style conditioning layers to convert the input style code to an intermediate illumination feature. The key difference is that the input of the style encoder is a histogram rather than an image. When inference, we could randomly select a histogram as the illumination input, at the same time, interpolation in the $z$ space between two histograms is also allowed. *The interpolation visualization video result can be seen on the project page.*

### B.3. RenderNet

The RenderNet is a variation of Pix2Pix [2]. As shown in Figure 1, the key difference is that we inject the style feature in the last three Upsample blocks, which includes the illumination information of the groundtruth image during training, thereby mitigating the effects of illumination changes.

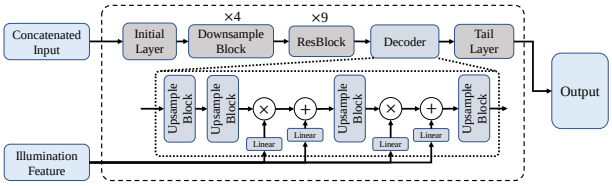


Figure 1. The architecture of RenderNet. We inject the illumination feature in the decoder.

### B.4. Discriminator

The discriminator we use is a multi-scale discriminator that differs from the vanilla multi-scale discriminator used in pix2pixHD [4]. While the vanilla discriminator operates on images of different scales, we use three discriminators: $D_1$, $D_2$, and $D_3$. $D_1$ works on panorama images, while $D_2$ and $D_3$ operate on perspective images obtained by randomly sampling from the input panorama using a perspective transformation, but at different scales. The two discriminators operate on perspective images because the distortion on the upper and lower bounds of the panorama is challenging for the convolution layer. Specifically, the field of view (FOV) of the sampled perspective images is 100. In our ablation study, all results use the same multi-scale discriminator. The input image size for $D_1$, $D_2$, and $D_3$ is $64\times256$, $64 \times 64$, and $32 \times 32$ respectively.

### B.5. Additional Training Details

The weight for L1 loss, L2 loss, KL loss, feature matching loss, perceptual loss, $\mathcal{L}_{\text{snop}}$ and GAN loss are 1, 10, 0.1, 10, 10, 1, 1 respectively when training. In volume rendering, we sample 100 points along each ray.

## C. Satellite and Panorama camera model

Actually, there are no given camera instincts in the original CVUSA and CVACT datasets, which only contain image pairs collected from Google Earth in the same location by GPS, we follow the assumptions in Shi *et al*. [3], which assumes that satellite images show the top of objects in an overhead view, which approximates parallel projection, while street-view panoramas capture scenes at ground level with a spherical equirectangular projection.

To describe a panoramic image with a 360-degree horizontal and 180-degree vertical field of view, we use the equirectangular projection and spherical coordinate system. To accomplish this, we assign the camera location as $\mathbf{o}$, and the width and height of the panorama image as $w$ and $h$, respectively. We use $x$ and $y$ as the pixel coordinates of the image pixel under consideration, and then we can use the following equations to determine the azimuthal and zenith angle $\theta$ and $\phi$, respectively:

$$\theta = \frac{2\pi x}{w}, \phi = \frac{\pi y}{h}$$

The equation allow us to determine the view direction $\mathbf{d}$ through any given image pixel.

We illustrate the orientation corresponding to the CVACT (align) dataset in Figure 2, where the same color indicates the same direction.

(a) satellite (b) panorama

Figure 2. Here is an example of an aligned satellite and ground panorama image pair from the training dataset. In the satellite image, the north direction is upward, while in the ground panorama image, the central column line represents the north direction. Both display the same red color. The central horizontal line in the panorama corresponds to the horizon.

## D. Discussion

### D.1. Urban scenes

The nadir satellite image can not see the vertical surfaces of tall buildings. Except for the issue of unseen vertical surfaces, two representative cases of tall buildings and transient objects (*e.g*., cars) will challenge our method, though we believe the geometric perspective would facilitate the task for urban scenes.

### D.2. Infinite region

We assume that **objects beyond the top view coverage in street view images only include the sky**. It is hard to find out which object (e.g. tree) lies outside of the satellite scene, for we have no real 3d shape to find it. Nevertheless, our assumption has shown clear effects, as evident from the ablation study (2:34-3:50 in the project page video).

### D.3. Assumption on horizontal ground planes

Sat2Density has the horizontal assumption as we did not know the camera location and world coordinate system. Another assumption we used is that the "world" is finite and limited by the satellite image. The used datasets follow these two assumptions and provide 1-to-1 paired data. Given by these facts, the movement of cameras is indeed on the ground plane with a constant height (*e.g*., 2m in our method and prior arts like [3]). From the perspective of view synthesis, such assumptions should work well, but it will inherently lead to inaccurate 3D scene geometry when the whole ground region of a scene is sloped. To resolve this problem, further studies could be explored with some new problems: (1) how to estimate the slope from a single orthogonally-rectified satellite image, and (2) how to define a proper world coordinate system and place the ground-view camera(s) in the world.

## References


[1] 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. 1

[2] 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. 1

[3] 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

[4] 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. 1