This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Cross-View Splatter: Feed-Forward View Synthesis with Georeferenced Images

Matias Turkulainen[1*] Akshay Krishnan[2] Filippo Aleotti[3] Mohamed Sayed[3]
Guillermo Garcia-Hernando[3] Juho Kannala[1,4] Arno Solin[1,5] Gabriel Brostow[6]
Daniyar Turmukhambetov[3]
[1]Aalto University [2]Georgia Tech [3]Niantic Spatial [4]University of Oulu [5]ELLIS Institute Finland [6]UCL

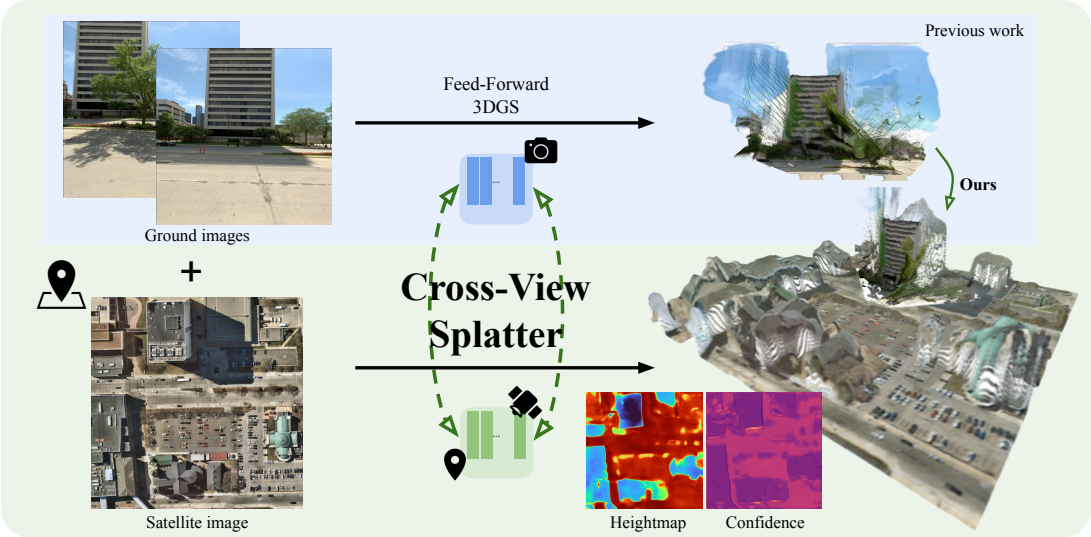


Figure 1. **Cross-View Splatter** is a feed-forward model that predicts Gaussian splats for GPS-tagged ground level images and corresponding orthorectified satellite imagery from mapping services. It predicts Gaussian splats for both ground level and bird's-eye views in a unified coordinate system and supports multiple input images with unknown 6DoF poses. Only the GPS location of the ground level images is required. Our model improves scene coverage and novel-view synthesis compared to using ground imagery alone.

Abstract

We present Cross-View Splatter, a feed-forward method that predicts pixel-aligned Gaussian splats for outdoor scenes captured at ground level AND by satellite. Faithful reconstructions require good camera coverage, but ground imagery is time-consuming and hard to capture at scale for large outdoor scenes. Fortunately, satellite imagery can provide a global geometric prior that is easy to access via public APIs. Cross-View Splatter fuses orthorectified satellite views with GPS-tagged ground photos to predict Gaussian splats in a unified 3D coordinate frame. By aligning ground and bird's-eye feature representations, our model improves scene coverage and novel-view synthesis, compared to ground imagery alone. We train on curated georeferenced datasets and paired satellite–terrain data, mined from open mapping services. We evaluate our method on a new benchmark for novel-view synthesis with georeferenced imagery allowing comparison to prior state-of-the-art methods. Our code and data preparation will be available at https://nianticspatial.github.io/cross-view-splatter/.

1. Introduction

Advances in 3D perception are transforming how visual systems model, interpret, and interact with the physical world. Generalized 3D reconstruction, in particular, has seen considerable success, with methods like [17, 38, 78, 80, 89, 97] allowing for accurate 3D point reconstruction using one or more images in a matter of seconds, with little to no prior information about camera calibrations. In addition, methods like [7, 35, 86, 90, 97] have extended feed-forward 3D reconstruction to novel-view synthesis by predicting Gaussian attributes for pixel-aligned points, which

*Work done as an intern at Niantic Spatial.

can be rendered via 3D Gaussian splatting [39], a desirable downstream task for 3D reconstruction. Despite this remarkable progress, a fundamental limitation still persists: these methods are designed, trained, and evaluated on ground level perspective imagery. This stems from the fact that widely used 3D foundation model training datasets consist predominantly of ground level calibrated views with aligned depth maps [3, 9, 26, 27, 45, 47, 62, 65, 92]. However, this reliance introduces significant scalability challenges, as modeling city-scale environments requires vast amounts of data that are both difficult to capture and computationally costly to process.

In contrast, satellite imagery provides a complementary perspective for outdoor captures. From a bird's-eye view (BEV), large-scale features such as roads and building footprints are more clearly delineated compared to ground-level imagery alone. High-quality satellite data is also readily available: mapping services such as Google Maps [28], Azure Maps [52], and Esri World Imagery [18] provide global surface imagery as fast, queryable tiled web maps (orthorectified satellite imagery). Our core insight is that, for geo-localized ground-level captures, tiled web-map satellite imagery provides a strong prior for global scene structure and geometry beyond street-view observations. Nonetheless, compared with ground-level or UAV imagery [14], satellite imagery remains challenging due to its coarser spatial resolution and variation in weather, illumination, and season, making it impractical for conventional structure-from-motion and 3D Gaussian splatting frameworks.

We therefore leverage satellite imagery in a learned manner to capture low-resolution scene geometry. We introduce Cross-View Splatter, a feed-forward method that uses both ground-level imagery and orthographic satellite imagery from freely available mapping services. Given one or more ground-level images with GPS and heading metadata, our model queries a corresponding BEV image. Ground and satellite images are encoded into a unified feature space via cross-attention [15] with a strong pre-trained 3D reconstruction model [35, 78]. This feature space is then used to regress 3D Gaussian splat attributes. We backproject splats from ground images using perspective projection and from satellite images using orthographic projection, then merge them into a single 3D scene representation for novel-view rendering.

However, orthoimagery poses a fundamental challenge. Orthorectification removes perspective effects and parallax, making 3D reconstruction from tiled web map imagery alone impossible for classical MVS methods [25, 41, 42, 60]. Sparse overlap between surfaces visible in satellite and ground views also makes feature-based 3D reconstruction [66] very challenging. To train our feed-forward method with suitable 3D data for ground and orthoimagery, we augment existing georeferenced ground-level datasets. We mine Digital Elevation Models (DEMs) and terrain data from public geological surveys [22, 52, 74] to create satellite-height map pairs, which provide a suitable 3D supervisory signal for satellite perspectives.

Contributions. To summarize, we provide the following contributions: (1) a novel feed-forward method capable of synthesizing Gaussian splats for both ground level perspective imagery and orthorectified satellite views, a first of its kind. We leverage satellite views alongside ground imagery to achieve greater coverage and a superior ability to extrapolate to novel views, an advantage over methods restricted to ground imagery alone. (2) Curated datasets with ground level imagery with georeferenced BEV satellite images and terrain height maps; and (3) a new challenging outdoor scenes benchmark of ground level images with aligned satellite views allowing comparison of state-of-the-art feed-forward 3D Gaussian splatting methods.

2. Related Work

Feed-forward reconstruction. Feed-forward reconstruction refers to learning based methods for reconstructing the geometry of a scene via a single forward pass of a neural network. In contrast to classical structure-from-motion (SfM) [8, 31, 55, 66] and multi-view stereo (MVS) [24, 29, 64] methods that reconstruct the scene with optimization schemes, feed-forward approaches are attractive for their speed and ability to generalize to challenging sparse inputs. The seminal work DUSt3R [81] introduced a ViT [15] encoder-decoder with self- and cross-attention that regresses two-view point maps in a canonical frame. Follow-up works [17, 38, 49, 50, 71, 78, 89] extended DUSt3R's two-view limited architecture to multi-view reconstruction. Methods like [2, 38, 78, 80, 83] also regress 6DoF camera poses relative to an arbitrary reference image, with [76] even using aerial perspective viewpoints. However, integrating orthoimagery remains an open and largely unexplored challenge.

Feed-forward novel-view synthesis. Learning based methods for view synthesis have also been proposed. Early works adopted multi-plane images (MPI) [98] as a scene representation for small baseline novel-view synthesis. They included single-view methods [73] and multi-view ones [69, 79, 98], and utilized CNN based architectures for feature extraction and 3D regression. With the advent of NeRF's [53], methods like [6, 16, 93] use multi-view architectures with NeRF-like volume rendering for view-synthesis.

The success of 3D Gaussian splatting [39] motivated a series of works, notably [5, 7, 32, 35, 68, 82, 86, 90], that regress per-pixel Gaussian attributes rendered to novel views via splatting-based rasterization. Methods such as

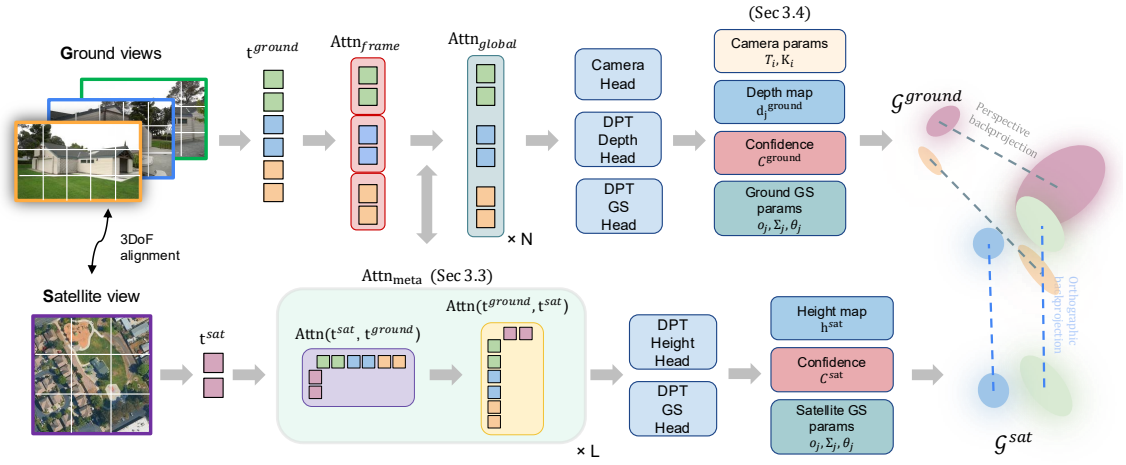


Figure 2. **Method overview**: Given geolocalized ground images and a single orthorectified satellite perspective, our model synthesizes 3D Gaussian splats in a shared coordinate frame. Ground views exchange information with satellite views within bidirectional cross-attention layers. Gaussians are predicted separately from ground and satellite branches, which are then combined into a unified coordinate frame. Although public elevation data is leveraged during training, only the satellite image and ground view(s) are necessary for inference.

NoPoSplat [90] and AnySplat [35] use strong pre-trained 3D foundation models [78, 81] as backbones, distilling their priors for view synthesis. This is a scalable approach, especially for datasets without ground-truth geometry [48, 98]. However, 3D foundation models are not trained on satellite imagery and therefore struggle in our cross-view setting. We adopt a hybrid strategy: we use pre-trained foundation models [35, 78] for ground-level geometry tasks, and fine-tune additional satellite-specific layers to align satellite and ground-level features.

Satellite images for 3D reconstruction. Prior work reconstructs city-scale 3D models from satellite imagery. Traditional methods [42, 60, 87] use MVS on non-rectified satellite and aerial images to recover geometry and BEV camera models (RPCs). More recent methods [25, 34, 36, 41] use Gaussian Splatting: SkySplat [34] is a multi-image feed-forward method for satellite images, Horizon-GS [36] jointly splats ground and aerial views, and Skyfall-GS [41] jointly optimizes splats with diffusion-based refinement. SkyEyes [25] follows a similar setup and reconstructs meshes from multi-view satellite imagery with surface-aligned Gaussians [30], refined in ground-level views via denoising [95]. A key limitation is reliance on multi-view off-nadir, non-orthorectified imagery, which is hard to acquire and typically unavailable from mapping services such as Google Maps. To address this, [33, 85] pair orthorectified satellite imagery with DEMs to predict satellite-aligned height maps, then convert them to voxel grids for ray-traced coarse ground-level depth. We also estimate satellite-aligned height maps at inference time, but unlike methods that warp this geometry for depth prediction, we directly render it into ground-level views using Gaussian splatting.

Satellite-to-ground view-synthesis. Generating ground-level novel views *solely* from satellite imagery has also been studied. Shi *et al*. [67] use relationships between orthographic views, height maps, and ground panoramas to warp satellite texture to ground level. Toker *et al*. [72] synthesize ground views with an intermediate polar transformation network. Xu *et al*. [87] reconstruct a scene mesh with MVS, project it to ground views, and texture it with a diffusion model. Sat2Density [58, 59] predicts a tri-plane representation [4] from top-down views that can be volume-rendered to known camera views, and unlike other methods can use a single satellite image. Inspired by these single-image methods, we adopt a Gaussian splat representation for top-down views that can be rasterized to novel views. Sky handling is also important: [59] regularizes sky regions in ground-level views to be opaque, whereas [58] generates sky color with StyleGAN-2 [37]. As in [59], we treat sky regions as part of the ground-level reconstruction model and promote sky Gaussians to be far away and opaque.

Generative satellite-to-ground methods have also been proposed. SatDreamer360 [94] predicts a tri-plane feature representation from BEV views, then synthesizes ground views by denoising a latent diffusion model [63] with ray-based cross-attention. Sat2GroundScape [88] reconstructs a mesh from satellite views and textures it with satellite-guided denoising. Sat2Scene [43] and Sat2Vid [44] instead synthesize a point cloud from satellite views, then denoise and texture it with a 3D diffusion model. Streetscapes [13] generates ground-level depth and semantics from Google Maps height maps and semantic labels, then uses them to condition a video diffusion model that generates ground-level videos of city blocks. In contrast, feed-forward approaches like ours synthesize only visible regions in ground and satellite views without hallucinating unobserved areas.

3. Model for Cross-View Splatter

3.1. Problem Formulation

Goal. Given an input sequence $(I_i^{\text{ground}})_{i=0}^N$ of ground level georeferenced imagery, we seek to reconstruct a 3D Gaussian splatting model in a feed-forward fashion with the aid of a *single* satellite perspective. Georeferenced imagery refers to images tagged with 3DoF pose information, specifically GPS latitude, longitude, and heading information (available in most devices with an IMU [1]), which enable retrieval of a corresponding top-down orthographic view I^{sat} with a known spatial resolution r^{sat}, with units of pixels per meter, from satellite mapping services [18, 28, 52].

Coordinate conventions. We construct a coordinate frame where I_0^{ground} serves as the reference, with the queried satellite image I^{sat} centered at this location. We consider I_0^{ground} as the reference frame and express images $I_{i>0}^{\text{ground}}$ relative to it. I_0^{ground} also defines the zero-altitude location. See Fig. 3 for more details. We use $i \in \{0, \dots, N-1\}$ to index images and $j \in \{0, \dots, M-1\}$ to index per-pixel attributes.

3.2. Preliminaries

Geometry transformer. We seek to encode ground level perspective imagery and orthoimagery into a similar feature space for 3D reconstruction. For ground level images, we adopt the VGGT [78] architecture as a feature extractor. Images $I_i^{\text{ground}} \in \mathbb{R}^{H \times W}$ are encoded using DINOv2 [54] to extract a sequence of patch tokens $t_i^{\text{img}} \in \mathbb{R}^{l_i \times d}$ with $d = 1024$ being the embedding dimension and $l_i = \frac{H \times W}{p^2}$ for a patch size of $p = 14$. A camera token $t_i^{\text{cam}} \in \mathbb{R}^{1 \times d}$ and four register tokens $t_i^{\text{reg}} \in \mathbb{R}^{1 \times d}$ [12] are prepended to each image token sequence. These concatenated token sequences $t_i^{\text{ground}} = (t_i^{\text{cam}}, t_i^{\text{reg}}, t_i^{\text{img}})$ are processed by a N-layer alternating-attention ViT transformer [15]. Each layer applies the attention mechanism to either its own image token sequence, referred to as frame attention $\text{Attn}_{\text{frame}}(t_i, t_i)$, or jointly with all image tokens referred to as global attention $\text{Attn}_{\text{global}}\left(\{t_i\}_{i=0}^N, \{t_i\}_{i=0}^N\right)$. Tokens t_i^{ground} are trained to regress 6DoF relative camera poses and intrinsics via the camera tokens t_i^{cam} and dense depth and point maps with DPT [61] heads.

Gaussian splatting. We express our 3D scene representation as a set $\mathcal{G} = \{(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, o_j, \boldsymbol{\theta}_j)\}_j$ of Gaussian distributions $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with opacities o_j and view-dependent colors $\boldsymbol{\theta}_j \in \mathcal{C}^{N_{\text{sh}}}$ represented via order $N_{\text{sh}} = 1$ spherical harmonics. We regress Gaussian parameters for both ground and satellite views, represented by $\mathcal{G}^{\text{ground}}$ and $\mathcal{G}^{\text{sat}}$ respectively. To render Gaussian splats we follow [39] and use depth-sorted alpha-blending of colors based on accumulated transmittance at each pixel location.

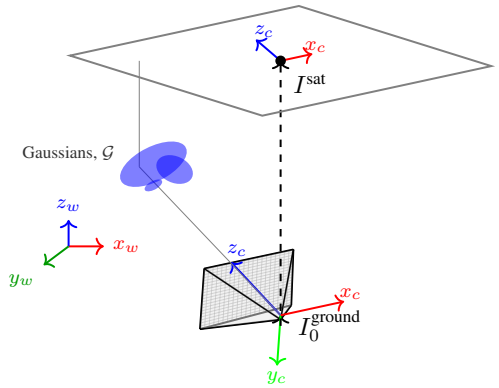


Figure 3. **Coordinate conventions**. We consider camera I_0^{ground} to define the origin of the world coordinates, *i.e.* the identity pose, as well as the spatial location of the BEV satellite image I^{sat}. The BEV I^{sat} frame is aligned with the heading of I_0^{ground} such that the camera look-at direction z_c is pointing up in the satellite view. Gaussian splats $\mathcal{G}$ are projected via perspective projection to ground views, and orthographic projection into BEV views. Other ground level images $I_{i>0}^{\text{ground}}$ are expressed relative to I_0^{ground}.

3.3. Joint Ground and Satellite Aware Transformer

The direct use of VGGT for orthorectified imagery is problematic: projections are not perspective, so they lack 6DoF poses and camera intrinsics and, more fundamentally, depth estimation is not usable without altitude information. Instead, we consider the transformation between ground and satellite views as 3DoF. We assume this transformation is known such that alignment as described in Fig. 3 holds. Similarly, we assume that the reference image I_0^{ground} defines the zero-altitude location. We thus formulate BEV geometry prediction as a height-map regression problem and estimating heights relative to the I_0^{ground} frame.

To achieve joint predictions with ground and satellite imagery, we inject cross-attention layers into the alternating-attention backbone of [78]. Specifically, we encode a BEV image $I^{\text{sat}} \in \mathbb{R}^{H_{\text{sat}} \times W_{\text{sat}}}$ as patch tokens $t^{\text{sat}} \in \mathbb{R}^{\frac{H_{\text{sat}} \times W_{\text{sat}}}{p^2} \times d}$ and introduce a bidirectional attention mechanism between ground tokens t_i^{ground} and t^{sat}, referred to as $\text{Attn}_{\text{meta}}$. This consists of two residual cross-attention layers that mix signals from satellite tokens t^{sat} and ground tokens t^{ground}, so

$$\text{Attn}_{\text{meta}}(t^{\text{sat}}, t^{\text{ground}}) = \mathcal{A}_2(t^{\text{sat}}, \mathcal{A}_1(t^{\text{ground}}, t^{\text{sat}}, t^{\text{sat}})), \quad (1)$$

where $\mathcal{A}(Q, K, V)$ is the multi-head attention mechanism [15, 75]. $\text{Attn}_{\text{meta}}$ is performed $L = 12$ times.

3.4. Geometry and Gaussian Splats Prediction

Ground views. Ground tokens t_i^{ground} flow through N-layers of alternating $\text{Attn}_{\text{frame}}$ and $\text{Attn}_{\text{global}}$ attention as well as L-layers of $\text{Attn}_{\text{meta}}$. Tokens t_i^{ground} are used to predict ground level depth maps and confidences with a DPT head:

$$d_j^{\text{ground}}, C_j^{\text{ground}} = \text{DPT}_{\text{depth}}(t_i^{\text{ground}}). \quad (2)$$

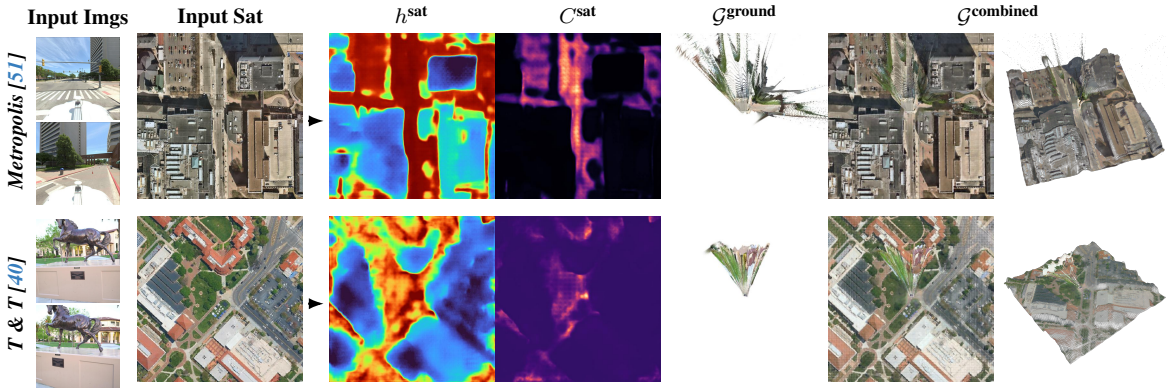


Figure 4. **Example reconstruction outputs** on scenes not seen during training. Left to right: input ground images, input satellite image, predicted height map, predicted height confidence (black: low, red: high), predicted ground Gaussians, predicted combined Gaussians.

Additionally, a camera-head is used to regress the 6DoF relative pose $\boldsymbol{T}_i$ and perspective camera intrinsics $\boldsymbol{K}_i$ for each input ground image using the camera tokens t_i^{cam}, which are used to define the means of ground Gaussians $\boldsymbol{\mu}_j^{\text{ground}} = \text{backproject}(d_j^{\text{ground}}, \hat{\boldsymbol{K}}_j, \hat{\boldsymbol{T}}_j)$. The remaining ground Gaussian parameters are predicted with a DPT head $\Sigma_j^{\text{ground}}, o_j^{\text{ground}}, \theta_j^{\text{ground}} = \text{DPT}_{\text{3DGS}}(t_i^{\text{ground}})$ that also has a skip connection from the input images to allow color information to propagate to the outputs [35].

Satellite views. Satellite tokens t^{sat} are used to regress a height map relative to I_0^{ground} and per-pixel confidence values C^{sat}, the latter of which is crucial for mitigating the impact of inaccurate or noisy ground-truth terrain height data:

$$h^{\text{sat}}, C^{\text{sat}} = \text{DPT}_{\text{height}}(t^{\text{sat}}). \tag{3}$$

Height map h^{sat} is transformed into 3D Gaussian locations μ_j^{sat} utilizing the known spatial resolution r^{sat} of the satellite image, so

$$\boldsymbol{\mu}_j^{\text{sat}} = \begin{pmatrix}\mu_x & \mu_y & \mu_z\end{pmatrix}^\top = \begin{pmatrix}\frac{u}{r^{\text{sat}}} & \frac{v}{r^{\text{sat}}} & h^{\text{sat}}(u,v)\end{pmatrix}^\top, \tag{4}$$

where u, v are pixel locations in I^{sat}. This conversion directly assumes an orthographic projection model for the satellite image. Similarly, the remaining satellite Gaussian $\mathcal{G}^{\text{sat}}$ attributes are regressed with a DPT head.

Scene scale normalization. We adopt the per-batch ℓ_2-scaling of the scene derived from backprojected depths, as established in [78, 81], to normalize depths, poses of ground imagery, and we also integrate $\mathbf{h}^{\text{sat}}$ and r^{sat} into the normalization scheme. See the Supplemental for further details. Notably, at inference time, all network predictions are expressed in the same normalized coordinate frame.

3.5. Losses

Ground views. We supervise our ground Gaussians $\mathcal{G}^{\text{ground}} = \{(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, o_j, \boldsymbol{\theta}_j)\}_j$ with the following losses. $\boldsymbol{\mu}_j^{\text{ground}}$ are regularized implicitly via depths d_j^{ground} with a confidence weighted depth loss

$$\mathcal{L}_{\text{depth}} = \sum_{j=1}^{M} \left(\left\| \hat{d}_j^{\text{ground}} - d_j^{\text{ground}} \right\|_2 - \alpha \log C_j \right). \tag{5}$$

During training, we utilize ground truth camera parameters, whereas at inference time we utilize predicted parameters. Camera poses and intrinsics are regularized with an L1 loss: $\mathcal{L}_{\text{cam}} = \|\hat{\boldsymbol{T}} - \boldsymbol{T}\|_1 + \|\hat{\boldsymbol{K}} - \boldsymbol{K}\|_1$. $\boldsymbol{\Sigma}_j$, i.e. the Gaussian sizes, are regularized with a depth consistency loss between predicted depths $\boldsymbol{d}_j^{\text{ground}}$ and Gaussian rendered depths, so

$$\mathcal{L}_{\text{const}} = \sum_{i=1}^{M} \left(\left\| d_j^{\text{ground}} - \hat{D}_{j,\text{3DGS}}^{\text{ground}} \right\| \right). \tag{6}$$

The Gaussian splat colors θ_j^{ground} and opacities o_j^{ground} are regularized with a mean-squared error and perceptual error [96] between ground-truth input images and Gaussian splats rendered from input image views, so this loss is

$$\mathcal{L}_{\text{RGB}}^{\text{ground}} = \sum_{i=1}^{N} \left(\|I_i - C_{i,\text{3DGS}}^{\text{ground}}\| + \gamma \cdot \text{LPIPS}(I_i, C_{i,\text{3DGS}}^{\text{ground}}) \right). \tag{7}$$

Satellite views. In a similar fashion, we regularize height predictions h^{sat} with a confidence weighted loss: $\mathcal{L}_{\text{height}} = \left\| \hat{h}^{\text{sat}} - h^{\text{sat}} \right\|_2 - \alpha \log C^{\text{sat}}$. Differently from ground views, for color supervision, we render satellite Gaussians to both input views as well as novel views, so

$$\mathcal{L}_{\text{RGB}}^{\text{sat}} = \sum_{i=1}^{N} \|I_i - C_{i,\text{3DGS}}^{\text{sat}}\| + \sum_{k=1}^{K} (\|I_{k,\text{nvs}} - C_{k,\text{3DGS}}^{\text{sat}}\|), \tag{8}$$

where I_{nvs} are selected as interpolating views between the bounds of the input images N.

Combined ground and BEV views. Lastly, we promote the synergy between the two Gaussian sets $\mathcal{G}^{\text{ground}}$ and $\mathcal{G}^{\text{sat}}$ by computing a loss on $\mathcal{G}^{\text{combined}} = \mathcal{G}^{\text{ground}} \cup \mathcal{G}^{\text{sat}}$, namely

$$\mathcal{L}_{\text{RGB}}^{\text{combined}} = \sum_{i=1}^{N} \|I_i - C_{i,\text{3DGS}}^{\text{combined}}\|. \tag{9}$$

BEV rendering. In Eq. (7) and Eq. (8), we regularize $\mathcal{G}^{\text{ground}}$ and $\mathcal{G}^{\text{sat}}$ with ground level views. To introduce a similar BEV constraint, we reproject Gaussian splats via orthographic projection onto the satellite plane. A key difference lies in the Gaussian size $\mathbf{\Sigma}_j$: while perspective Gaussian splatting scales $\mathbf{\Sigma}_j$ proportionally to the inverse z-depth (distance between camera $\mathbf{T}_i$ and Gaussian mean $\boldsymbol{\mu}_j$), orthographic projection omits this depth-dependent scaling. Therefore, when transforming $\mathcal{G}^{\text{ground}}$ and $\mathcal{G}^{\text{sat}}$ into the BEV perspective, we directly project normalized world coordinates into satellite pixel space using the spatial resolution r^{sat} factor of the satellite image. We compute a combined Gaussian rendering loss in the satellite view, so

$$\mathcal{L}_{BEV} = \|I^{\text{sat}} - C_{\text{3DGS}}^{\text{combined}}\|, \tag{10}$$

where $C_{\text{3DGS}}^{\text{combined}}$ denotes renders from combined Gaussians $\mathcal{G}^{\text{combined}} = \mathcal{G}^{\text{ground}} \cup \mathcal{G}^{\text{sat}}$.

Sky regularization. Sky pixels lack reliable depth cues and often produce erroneous depth estimates. To handle such regions, we identify sky pixels using an off-the-shelf segmentation model [46, 77, 100] and penalize implausibly close depth estimates using

$$\mathcal{L}_{\text{sky_depth}} = \sum_{j=0}^{M} \boldsymbol{M}_j \cdot \text{ReLU}(\tau - d_j^{\text{ground}}), \tag{11}$$

where τ is a threshold distance and

$$\boldsymbol{M}_j = \begin{cases} 1 & \text{if sky} \\ 0 & \text{otherwise.} \end{cases}$$

We also observe that explicitly promoting opaqueness of the sky pixels is necessary, so

$$\mathcal{L}_{\text{sky_alpha}} = \sum_{j=0}^{M} \boldsymbol{M}_j \cdot \|1 - o_j\|_1. \tag{12}$$

We denote the total sky loss with $\mathcal{L}_{\text{sky}} = \mathcal{L}_{\text{sky_depth}} + \mathcal{L}_{\text{sky_alpha}}$.

3.6. Georeferenced Data Curation

To train our model, we collected a set of geolocalizable ground level datasets, then augment those with BEV satellite images and terrain height data. For orthoimages, we use tiled web map providers Google Maps [28], Azure Maps [52], and Esri World Imagery [18]. For access, we used [19–21] and query views at ground level GPS locations at a sampling density of 2 pixels per meter, at resolution 512×512. For terrain data, we leverage government and public lidar data, namely Geological Survey data [74] and various data from [22]. We illustrate training samples in Fig. 5 and give a breakdown of datasets in Table 1. We give more examples and details in the Supplemental. Due to licensing restrictions (*e.g.* Google Maps, Azure Maps) we host only our additional data, and will provide code to query for the satellite images and reproduce the training datasets.

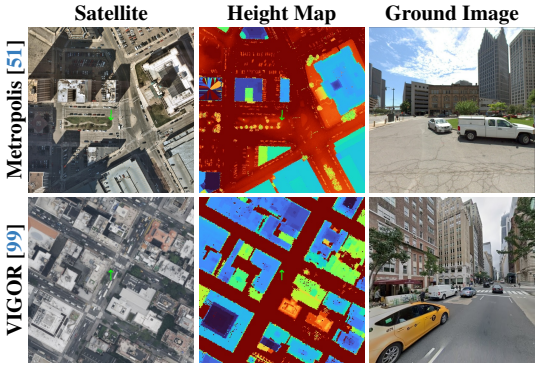


Figure 5. Illustration of training samples from our BEV augmented data. Ground 3DoF pose is denoted with green ↑ arrow.

Table 1. **Our training data** consists of georeferenced outdoor images that include 3DoF pose information, satellite imagery, terrain height maps, and/or ground level depth data. Our driving scenes are sourced from Metropolis and VIGOR. VIGOR contains panoramas, we create perspective cutouts with 90° FoV. * we generate pseudo-ground truth depths using UniK3D [57].

Name	Satellite	Terrain Maps	Depth Maps	# scenes	Static scenes?
Metropolis [51]	✓	✓	✓	75	-
VIGOR [99]	✓	✓	*	52K	✓
MapFree [1]	-	-	✓	655	✓
VKITTI2 [3, 23]	-	-	✓	21	-
DL3DV[47]	-	-	✓	10K	✓

4. Experiments

We evaluate Cross-View Splatter on novel view synthesis.

Datasets. We introduce a new task: *view synthesis with geolocalized outdoor images*. Because no benchmark currently exists for this setting, we construct one by augmenting the well known Tanks and Temples dataset [40] and selected outdoor scenes from the DL3DV-Benchmark [47]. Specifically, we manually align and scale the COLMAP [66] reconstructions of 10 outdoor scenes from Tanks and Temples and 40 scenes from DL3DV to satellite imagery. This alignment process is detailed in the Supplemental. We then evaluate our method and competing baselines in a sparse view setting with varying numbers of input views.

Training Data. We train with a combination of ground-

Table 2. **Outdoor Tanks and Temples sparse-view synthesis.** Metrics are averaged over 10 scenes. For Cross-View Splatter, we report ground-only, terrain-only, and combined (ground+terrain) reconstructions. Methods marked with * use ground-truth intrinsics. Methods marked with * require multi-view input and were given one additional adjacent frame during testing. Sat2Density† takes a single satellite image stylized with one context image (see Fig. 6).

Method		GT Pose?	1 context view			2 context views			3 context views		
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Splatfacto		✓	-	-	-	11.53	0.2611	0.6436	11.72	0.2888	0.6267
MVSplat		✓	-	-	-	6.93	0.1252	0.6997	7.58	0.1631	0.6941
DepthSplat		✓	-	-	-	9.61	0.3146	0.6077	10.72	*0.3557*	0.5873
NoPoSplat		*	6.43*	0.1062*	0.7040*	8.97	0.2197	0.6830	8.82	0.2359	0.6825
Long-LRM		-	8.53	0.3392	0.7054	8.53	0.3392	0.7054	10.54	0.3253	0.6477
AnySplat		-	7.48	*0.3572*	*0.6482*	*9.85*	0.3483	**0.5773**	*10.93*	*0.3775*	**0.5331**
Sat2Density†		✓	*8.81*	0.3557	0.8172	8.90	*0.3507*	0.8097	8.85	0.3508	0.8037
Ours	*Combined*	-	**11.13**	**0.3764**	**0.6286**	**11.67**	**0.3725**	*0.5984*	**12.00**	**0.3855**	*0.5699*
	Ground	-	8.92	0.3621	0.6066	9.94	0.3615	0.5877	10.61	0.3763	0.5631
	Terrain	-	8.39	0.3783	0.6257	9.82	0.4341	0.7474	9.63	0.4301	0.7472

Table 3. **Outdoor DL3DV sparse-view synthesis results.** Results are averaged over 40 scenes.

Method		GT Pose?	1 context view			2 context views			3 context views		
			PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Splatfacto		✓	-	-	-	13.46	0.2962	0.6158	13.61	0.3018	0.6026
MVSplat		✓	-	-	-	6.27	0.0413	0.7174	6.29	0.0474	0.7158
DepthSplat		✓	-	-	-	8.58	0.1569	0.6774	9.04	0.1817	0.6761
NoPoSplat		*	6.89*	0.0669*	0.7019*	*11.01*	0.2670	0.6665	*11.10*	0.2731	0.6687
Long-LRM		-	4.78	**0.3153**	0.7196	9.74	0.2842	0.6813	10.93	0.2890	0.6149
AnySplat		-	*8.37*	0.2639	*0.6498*	10.37	**0.3014**	**0.5702**	10.88	*0.3122*	**0.5557**
Ours	*Combined*	-	**11.33**	*0.2741*	**0.6307**	**12.10**	*0.2976*	*0.5940*	**12.61**	**0.3204**	*0.5683*
	Ground	-	9.00	0.2592	0.6191	10.05	0.2878	0.5842	10.65	0.3103	0.5606
	Terrain	-	8.24	0.2790	0.6884	8.41	0.2801	0.6932	8.30	0.2834	0.6928

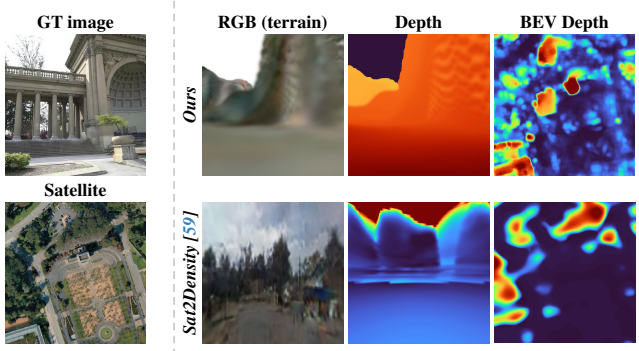


Figure 6. **Satellite-to-ground qualitative results. Column 1:** target ground image and satellite image input. **Columns (2-4): Predictions.** *(Top)* our $\mathcal{G}^{\text{sat}}$ rendered to ground-view RGB, ground-view depth, and BEV depth. *(Bottom)* **Sat2Density** [59] with RGB and volume-rendered ground/BEV depths from predicted density. Our method produces sharper, more accurate depth maps. The scene is challenging for Sat2Density [59], as it was predominantly trained on US country-side images.

view only datasets and those that we augment with satellite imagery, see Table 1 and Sec 3.6. Note that for DL3DV[47], we remove benchmark scenes from our training data.

Baselines. We consider various feed-forward novel-view methods. We compare against a) MVSplat [7] and b) DepthSplat [86] for cost-volume based multi-view 3DGS methods c) AnySplat [35], a feed-forward method initialized by the [78] model weights, d) NoPoSplat [90], a similar feed-forward method initialized by [81] weights, and the e) Splatfacto [70] method, our baseline that performs per-scene optimization using input views taking 5+ minutes to optimize each scene. We also provide qualitative results comparing satellite-to-ground-only view-synthesis using Sat2Density [59].

As DL3DV-10K includes DL3DV-Benchmark scenes and NoPoSplat trains on the entire set, in Table 3 we evaluate the RealEstate10K [98] variant of NoPoSplat. We do not compare against FLARE [97], as their training data contains scenes of DL3DV-Benchmark and scenes of Tanks and Temples within the Megadepth dataset. For fairness, we center-crop input and target images to square images and rescale to each method's training resolution.

View Selection. First, we compute frame overlaps based on the number of co-visible sparse points recovered by Colmap across every view pair. For context views, we start by selecting the first index in each scene and then we select each subsequent required context view by optimizing for a target overlap. For each combination of context views, we select four target views, each at varying overlap targets to the context views. See the Supplemental for more details.

Evaluation metrics. We report PSNR, SSIM [84], and LPIPS (VGG-net) [96] metrics. For our method, we report metrics for ground level Gaussians (*ground*), satellite-to-ground level Gaussians (*terrain*), and combined ground and satellite Gaussians (*combined* (Ours)).

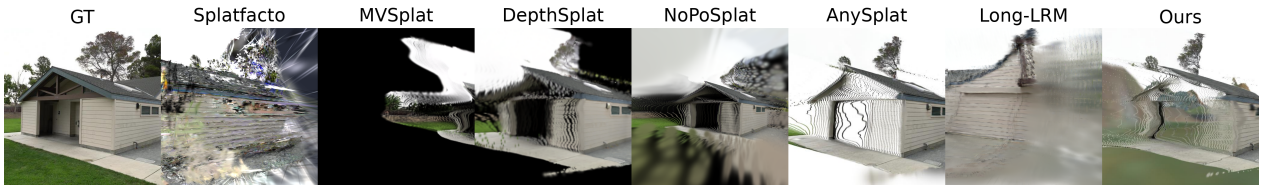


Figure 7. **Qualitative results.** We show results for a target image from the Outdoor Tanks and Temples dataset under the sparse-view synthesis setting. Our method extrapolates into unobserved regions by leveraging visual cues from the corresponding satellite image.

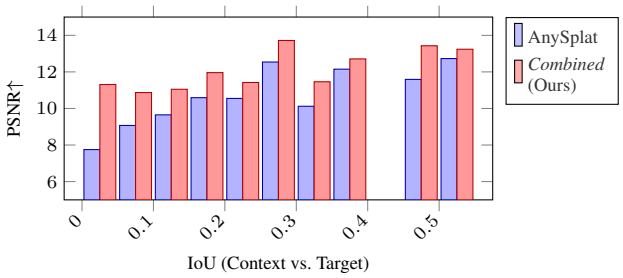


Figure 8. **Stratified evaluation**. Bucketed PSNR performance (5% bins) vs. image overlap on our geolocalized Tanks & Temples dataset.

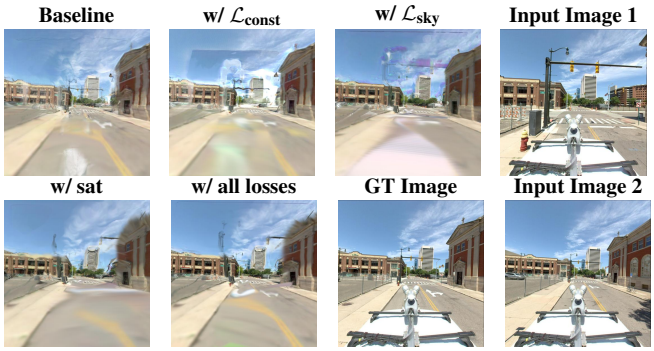


Figure 9. **Qualitative ablation** of model with 2 input images.

Implementation details. We use PyTorch [56] with `gsplat (v1.5)` [91] for Gaussian rasterization. We initialize our model from AnySplat [35]. We train for 4 days on 2×-A100 GPUs with a batch size of 10, using FlashAttention-v2 [10, 11] and mixed-precision. We resize satellite images and terrain heights such that their spatial extent is 244 meters. Our input resolution is 518×518.

4.1. Results and Ablations

We report novel view synthesis results in Tables 2 and 3. Our method matches or outperforms baselines, with the satellite head (*Combined*) improving scores across all context coverage levels. Although overall PSNR values are low, this reflects the challenging setup—input–target pairs have low overlap (IoU $\approx$ 0.05–0.5), making many cases difficult. Cost-volume baselines like MVSplat and DepthSplat require multiple inputs and tend to fail under 1-view or small-baseline conditions, where our satellite-conditioned model performs more robustly. Qualitative comparisons are shown in Fig. 7 and Fig. 6.

In Fig. 8, we stratify Tanks and Temples by context–target IoU (co-visible COLMAP features at 5% bins) and plot PSNR. The satellite model outperforms the baseline across bins, with the largest gains at low overlap ($\leq$ 0.15 IoU), indicating stronger extrapolation to novel views. Ablations in Tab. 4 and Fig. 9 show that consistency and sky regularization improve VGGT w/ 3DGS, while VGGT w/ 3DGS w/ SAT performs best, especially with $\mathcal{L}_{\text{RGB}}^{\text{sat}}$, likely due to better BEV coverage of occluded and unseen regions.

Table 4. **Quantitative ablation** on the Metropolis dataset. Metrics are averaged over 36 test scenes, covering 2 input and 2 interpolated novel-views. Our satellite variant achieves superior results due to its larger BEV coverage, allowing for extrapolation into occluded and unseen regions in large baseline driving scenes.

Method	Ground PSNR	Terrain PSNR	Combined PSNR
VGGT w/ 3DGS:			
$\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{RGB}}^{\text{ground}}$	15.26	-	-
$+\mathcal{L}_{\text{const}}$	16.99	-	-
$+\mathcal{L}_{\text{sky}}$	17.10	-	-
VGGT w/ 3DGS w/ SAT:			
$\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{const}} + \mathcal{L}_{\text{sky}} + \mathcal{L}_{\text{RGB}}^{\text{combined}}$	16.99	5.24	17.17
$+\mathcal{L}_{\text{RGB}}^{\text{ground}}$	16.61	5.36	16.87
$+\mathcal{L}_{\text{RGB}}^{\text{sat}}$	17.59	12.25	18.63

5. Conclusions

We introduced Cross-View Splatter, a novel feed-forward method that predicts Gaussian splats for ground *and* satellite images. Unlike previous approaches, our BEV Gaussians increase scene coverage and improves view-synthesis in difficult outdoor scenes. To train our model with the new cross-view setting, we curated a collection of existing datasets and supplemented them with satellite views and height maps. We also introduce a new benchmark for novel-view synthesis with the aid of satellite images using public outdoor scenes of Tanks and Temples and DL3DV.

6. Acknowledgments

We thank Zawar Qureshi and Jakub Powierza for compute infrastructure support and Alan Paul for help in generating terrain data. MT, JK, and AS acknowledge funding from the Research Council of Finland (362408, 339730).

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, 2022. 4, 6

[2] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2

[3] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 2, 6

[4] 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, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 3

[5] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024. 2

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *arXiv preprint arXiv:2103.15595*, 2021. 2

[7] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, 2024. 1, 2, 7

[8] David J. Crandall, Andrew Owens, Noah Snavely, and Daniel P. Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *PAMI*, 2013. 2

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[10] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024. 8

[11] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS*, 2022. 8

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 4

[13] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *SIGGRAPH*, 2024. 3

[14] Oussema Dhaouadi, Riccardo Marin, Johannes Michael Meier, Jacques Kaiser, and Daniel Cremers. Ortholoc: UAV 6-dof localization and calibration using orthographic geodata. In *NeurIPS Datasets and Benchmarks Track*, 2025. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 2, 4

[16] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. *CVPR*, 2023. 2

[17] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 1, 2

[18] Esri. Esri World Imagery. https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9. Accessed: 2025-10-05. 2, 4, 6

[19] Florian Fervers. Tiled Web Maps. https://github.com/fferflo/tiledwebmaps. Accessed: 2025-10-04. 6

[20] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *CVPR*, 2023.

[21] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Statewide visual geolocalization in the wild. In *ECCV*, 2024. 6

[22] Flai. Collection of open nation-scale lidar datasets. https://registry.opendata.aws/open-lidar-data. Accessed: 2025-10-19. 2, 6

[23] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 6

[24] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. 2

[25] Zhiyuan Gao, Wenbin Teng, Gonglin Chen, Jinsen Wu, Ningli Xu, Rongjun Qin, Andrew Feng, and Yajie Zhao. Skyeyes: Ground roaming using aerial view images. *arXiv preprint arXiv:2409.16685*, 2024. 2, 3

[26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 2

[27] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012. 2

[28] Google. Google Maps Platform Documentation. https://developers.google.com/maps/documentation. Accessed: 2025-10-04. 2, 4, 6

[29] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020. 2

[30] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 2024. 3

[31] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2

[32] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim.

Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024. 2

[33] Xiaomou Hou, Wanshui Gan, and Naoto Yokoya. Enhancing monocular height estimation from aerial images with street-view images. *arXiv preprint arXiv:2311.02121*, 2023. 3

[34] Xuejun Huang, Xinyi Liu, Yi Wan, Zhi Zheng, Bin Zhang, Mingtao Xiong, Yingying Pei, and Yongjun Zhang. Skysplat: Generalizable 3d gaussian splatting from multi-temporal sparse satellite images. *arXiv preprint arXiv:2508.09479*, 2025. 3

[35] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025. 1, 2, 3, 5, 7, 8

[36] Lihan Jiang, Kerui Ren, Mulin Yu, Linning Xu, Junting Dong, Tao Lu, Feng Zhao, Dahua Lin, and Bo Dai. Horizon-gs: Unified 3d gaussian splatting for large-scale aerial-to-ground scenes. In *CVPR*, 2025. 3

[37] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3

[38] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 1, 2

[39] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 2, 4

[40] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *TOG*, 2017. 5, 6

[41] Jie-Ying Lee, Yi-Ruei Liu, Shr-Ruei Tsai, Wei-Cheng Chang, Chung-Ho Wu, Jiewen Chan, Zhenjun Zhao, Chieh Hubert Lin, and Yu-Lun Liu. Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. *arXiv preprint arXiv:2510.15869*, 2025. 2, 3

[42] Matthew J. Leotta, Cheng Long, Bastien Jacquet, Michael Zins, Daniel Lipsa, Jizhe Shan, Boyan Xu, Zhaoyu Li, Xun Zhang, Shih-Fu Chang, Misu Purri, Jia Xue, and Kristin Dana. Urban Semantic 3D Reconstruction from Multiview Satellite Imagery. In *CVPRW*, 2019. 2, 3

[43] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R. Oswald. Sat2Scene: 3D urban scene generation from satellite images with diffusion. In *CVPR*, 2024. 3

[44] 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 *ICCV*, 2021. 3

[45] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 2

[46] Orly Liba, Longqi Cai, Yun-Ta Tsai, Elad Eban, Yair Movshovitz-Attias, Yael Pritch, Huizhong Chen, and Jonathan T Barron. Sky optimization: Semantically aware image processing of skies in low-light photography. In *CVPRW*, 2020. 6

[47] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 2, 6, 7

[48] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 3

[49] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *CVPR*, 2025. 2

[50] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025. 2

[51] Mapillary. Mapillary Metropolis Dataset. https://www.mapillary.com/dataset/metropolis. Accessed: 2025-10-18. 5, 6

[52] Microsoft. Azure Maps. https://azure.microsoft.com/en-us/products/azure-maps. Accessed: 2025-10-04. 2, 4, 6

[53] 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 *ECCV*, 2020. 2

[54] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[55] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *ECCV*, 2024. 2

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Nataly Gimelshein, Luca Antiga, Alban Kopf, Federico Metta, Allan Chiley, Brian Stwalley, Sheng Huang, Jiawan Jiang, Yehezkel Chen, Peng Zeng, Xiaobing Li, James Yu, Teteya Li, Andrey Kuchaiev, Kartik Ren, Houdong Zhang, Yanghan Shi, Jani Singh, Wenzhe Yu, Gregory Pierson, Soumith Chintala, and Yann LeCun. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 8

[57] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3d estimation. In *CVPR*, 2025. 6

[58] Ming Qian, Bin Tan, Qiuyu Wang, Xianwei Zheng, Han-

jiang Xiong, Gui-Song Xia, Yujun Shen, and Nan Xue. Seeing through satellite images at street views. *arXiv preprint arXiv:2505.17001*, 2025. 3

[59] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *ICCV*, 2023. 3, 7

[60] Rongjun Qin. 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*, 2016. 2, 3

[61] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 4

[62] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 2

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 3

[64] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2

[65] Thomas Schöps, Torsten Sattler, Christian Häne, and Marc Pollefeys. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 2

[66] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 6

[67] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *PAMI*, 2022. 3

[68] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2

[69] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. *CVPR*, 2019. 2

[70] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023. 7

[71] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander G. Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, 2025. 2

[72] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization. In *CVPR*, 2021. 3

[73] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 2

[74] U.S. Geological Survey. USGS Lidar Explorer Map. https://apps.nationalmap.gov/lidar-explorer. Accessed: 2025-10-19. 2, 6

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[76] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *CVPR*, 2025. 2

[77] Jianyuan Wang. Skyseg. https://huggingface.co/JianyuanWang/skyseg. Accessed: 2025-08-10. 6

[78] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 3, 4, 5, 7

[79] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[80] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 1, 2

[81] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 3, 5, 7

[82] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free-view synthesis of indoor scenes. *arXiv preprint arXiv:2405.17958*, 2024. 2

[83] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 2

[84] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 7

[85] Scott Workman and Hunter Blanton. Augmenting depth estimation with geospatial context. In *ICCV*, 2021. 3

[86] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, 2025. 1, 2, 7

[87] Ningli Xu and Rongjun Qin. Geospecific view generation – geometry-context aware high-resolution ground view inference from satellite views. In *ECCV*, 2024. 3

[88] Ningli Xu and Rongjun Qin. Satellite to GroundScape: Large-scale consistent ground view generation from satellite views. In *CVPR*, 2025. 3

[89] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *CVPR*, 2025. 1, 2

[90] Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025. 1, 2, 3, 7

[91] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *JMLR*, 2025. 8

[92] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2

[93] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[94] Xianghui Ze, Beiyi Zhu, Zhenbo Song, Jianfeng Lu, and Yujiao Shi. SatDreamer360: Geometry consistent street-view video generation from satellite imagery. *arXiv preprint arXiv:2506.00600*, 2025. 3

[95] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, 2023. 3

[96] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 7

[97] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *CVPR*, 2025. 1, 7

[98] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 2, 3, 7

[99] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *CVPR*, 2021. 6

[100] Xiong Zhu. Sky-segmentation-and-post-processing. https://github.com/xiongzhu666/Sky-Segmentation-and-Post-processing. Accessed: 2025-08-10. 6