AerialMegaDepth: Learning Aerial-Ground Reconstruction and View Synthesis

Khiem Vuong Anurag Ghosh
Deva Ramanan* Srinivasa Narasimhan* Shubham Tulsiani*

Carnegie Mellon University
https://aerial-megadepth.github.io

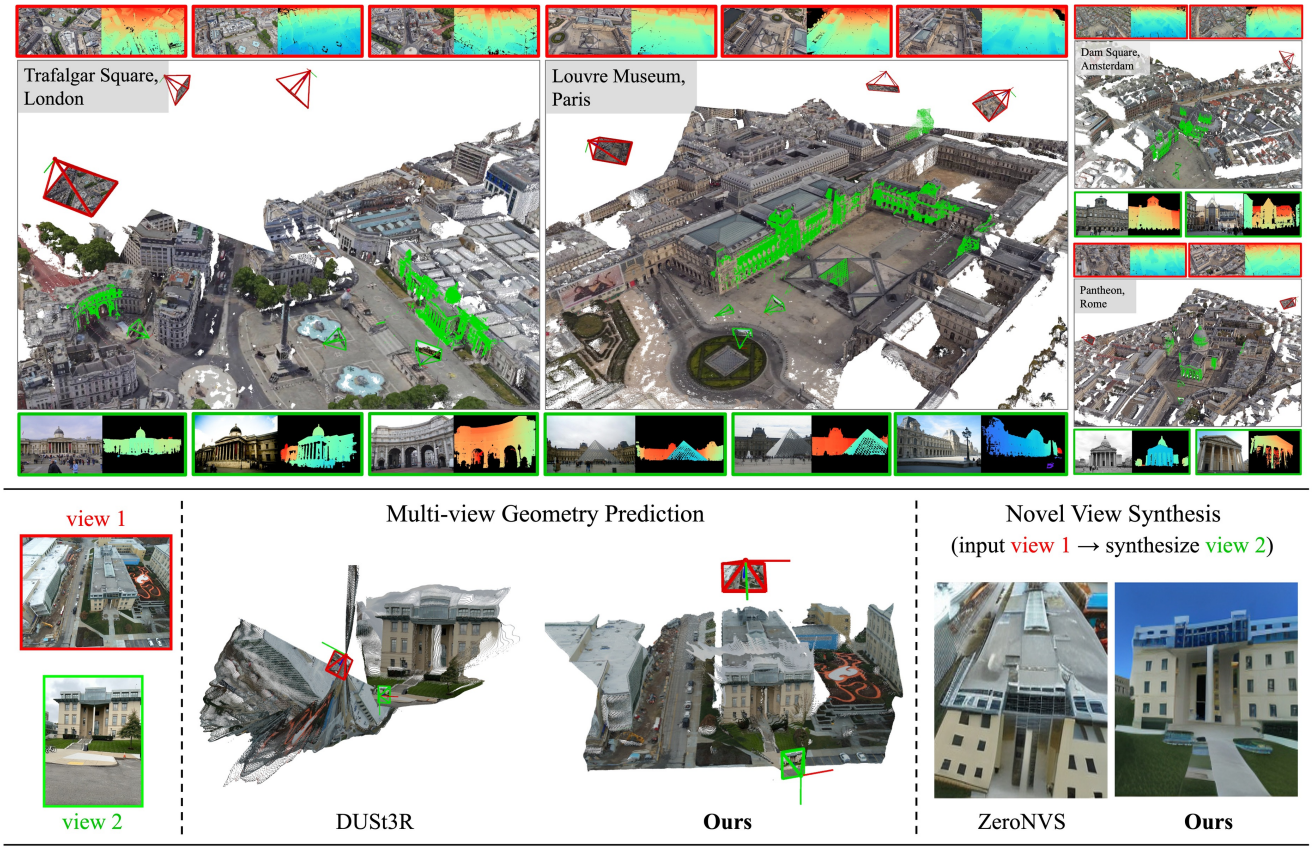

Figure 1. **First row:** Examples of our generated cross-view (aerial-ground) geometry data, including co-registered pseudo-synthetic (i.e., mesh-rendered) aerial and real ground-level images, with corresponding depth maps, point clouds, and camera intrinsics/extrinsics in a unified coordinate system, for a variety of scenes. **Second row:** Leveraging such data curated over 137 landmarks and 132K geo-registered images, we show significant improvements in learning-based methods on **real unseen** ground-aerial scenarios across two representative tasks: 1) multi-view geometry prediction using DUSt3R [64] finetuned on our data, and 2) novel view synthesis from a single image conditioned on a target pose by fine-tuning ZeroNVS [45] that was originally trained on MegaScenes [60].

Abstract

We explore the task of geometric reconstruction of images captured from a mixture of ground and aerial views. Current state-of-the-art learning-based approaches fail to handle the extreme viewpoint variation between aerial-ground image pairs. Our hypothesis is that the lack of high-quality, co-registered aerial-ground datasets for training is a key reason for this failure. Such data is difficult to assemble precisely because it is difficult to reconstruct in a scalable way. To overcome this challenge, we propose a scalable framework combining pseudo-synthetic renderings from 3D city-wide meshes (e.g., Google Earth) with real, ground-level

* denotes equal contribution/advising

crowd-sourced images (e.g., MegaDepth [29]). The pseudo-synthetic data simulates a wide range of aerial viewpoints, while the real, crowd-sourced images help improve visual fidelity for ground-level images where mesh-based renderings lack sufficient detail, effectively bridging the domain gap between real images and pseudo-synthetic renderings. Using this hybrid dataset, we fine-tune several state-of-the-art algorithms and achieve significant improvements on real-world, zero-shot aerial-ground tasks. For example, we observe that baseline DUSt3R [64] localizes fewer than 5% of aerial-ground pairs within 5 degrees of camera rotation error, while fine-tuning with our data raises accuracy to nearly 56%, addressing a major failure point in handling large viewpoint changes. Beyond camera estimation and scene reconstruction, our dataset also improves performance on downstream tasks like novel-view synthesis in challenging aerial-ground scenarios, demonstrating the practical value of our approach in real-world applications.

1. Introduction

The ability to register, reconstruct, or generally reason about multi-view images has been a cornerstone task in computer vision. While classical pipelines [50, 51] leveraged hand-designed features [3, 35, 44] and matching mechanisms, their subsequent incarnations [7, 39, 63] have incorporated several learning-based components *e.g.* learned features [13, 32] or learned "doppelganger" detectors [8]. More recently, approaches have departed from this classical pipeline to directly learn multi-view tasks such as 2D correspondences [33, 46, 58], camera estimation [30, 62, 72], pointmap prediction [26, 64] and novel-view synthesis [34, 45, 60] in an end-to-end manner. This shift towards learning-based components and approaches has led to impressive progress, particularly in challenging scenarios, e.g., sparsely sampled input, or varying illumination.

Much of this progress has been fueled by large crowd-sourced image collections like MegaDepth [29], which provides accurate 3D reconstructions built using structure-from-motion (SfM) from thousands of tourist-uploaded images at various landmarks. This 3D data offers valuable supervision for geometric tasks and has been transformative for multi-view learning algorithms. However, because these images are primarily captured by tourists, they mostly cover ground-level viewpoints, and in some cases, aerial viewpoints, but rarely both. As a result, methods like DUSt3R [64] and MASt3R [26], though trained on diverse "in-the-wild" datasets including MegaDepth, struggle with large viewpoint changes between handheld (ground-level) and drone-mounted (aerial) views, as shown in Figure 1. For example, in our evaluation, pre-trained DUSt3R achieves only a $\sim$5% success rate in registering cameras (under 5° rotational error) from ground-aerial pairs.

Our central hypothesis is that this limitation stems from the lack of training data that contains co-registered ground-aerial image pairs. While independent ground and aerial camera poses are easy to obtain, merging them into a unified coordinate system often requires specialized sensors or manual effort, limiting scalability. To address this, we propose a flexible and scalable data generation framework leveraging geospatial platforms like Google Earth, which render 3D textured meshes of cities and landmarks – providing another vast data source. We refer to these mesh renderings as *pseudo-synthetic* since they are rendered from 3D meshes of actual landmarks textured with real photos. We construct pseudo-synthetic data by rendering aerial-ground viewpoints at varying altitudes from these textured meshes. While using these images alone show promising improvements, such mesh renderings from ground-level viewpoints are not as photorealistic, with a domain gap due to differences in lighting, texture, and other visual details compared to real images. To mitigate this, we propose to co-register abundantly available real ground images (e.g., from MegaDepth [29]) with the pseudo-synthetic images within the same coordinate frame. The pseudo-synthetic data, captured at varying altitudes, simulates a wide range of aerial viewpoints, while the real, crowd-sourced images help improve visual fidelity, especially for ground-level images where mesh renderings lack details. We call this hybrid dataset *AerialMegaDepth*, and Figure 1 shows some landmarks from our data, with 132,137 co-registered real and pseudo-synthetic images across 137 scenes.

From this dataset, we generate over 1.5 million aerial-ground image pairs and fine-tune several state-of-the-art reconstruction algorithms, showing significant improvements on real-world mixed-altitude imagery (see Figure 1 for examples). Quantitatively, fine-tuning 3D prediction models like DUSt3R [64] and MASt3R [26] improves the camera registration success rate (with rotation error under 5°) from just 5% to nearly 56%. In addition, our dataset also improves novel-view synthesis in challenging aerial-ground scenarios (see Figure 1). Finally, we emphasize that our framework is flexible and scalable, applicable not only to MegaDepth but also to other crowd-sourced datasets [6, 22, 36, 60] and geospatial platforms [9, 12, 24], making it possible to leverage a nearly unlimited source of data to learn aerial-ground 3D reconstruction.

2. Related Works

Datasets for Multiview 3D and Geometry. Several datasets released in the last decade have pushed the frontier of multiview 3D geometry. Internet-sourced datasets of landmarks like MegaDepth [29] and IMC-PT [23] mostly have ground-level views. But still, these datasets, whose ground-truth was built using SfM [50], have spurred monocular depth estimation [5, 68], multiview stereo [26, 64], learned feature matching [19, 46, 58, 76], learned pose

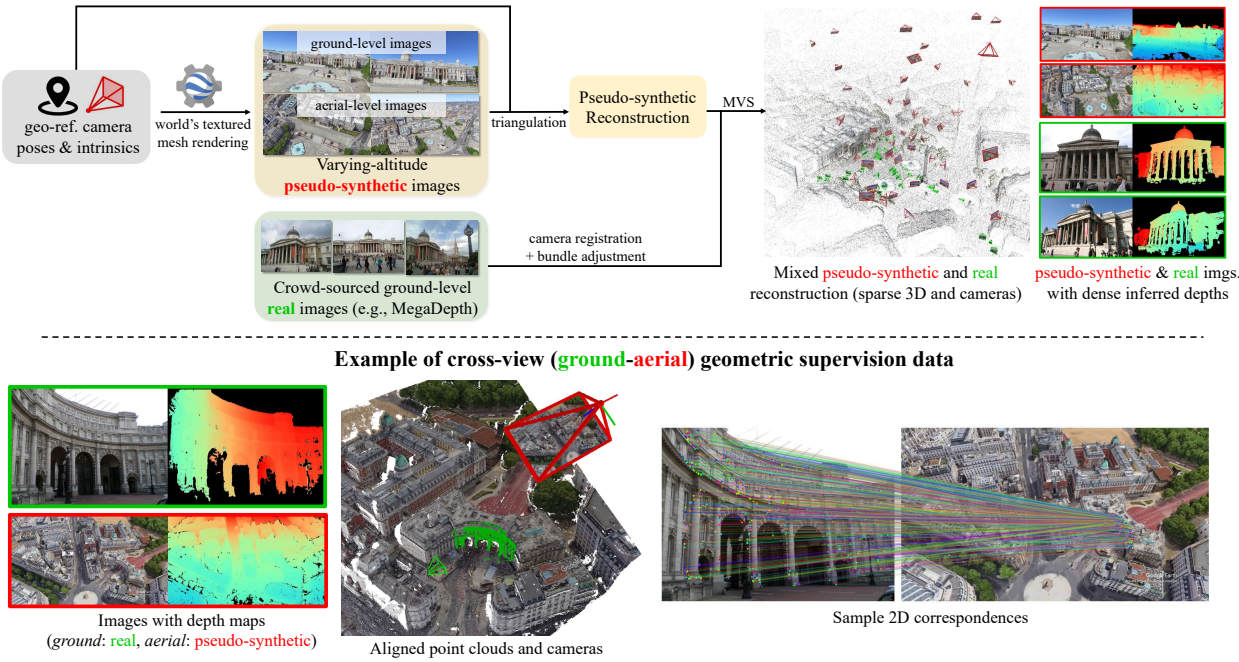

Figure 2. **Overview of the data generation framework.** To address the challenges of ground-aerial camera registration and novel-view synthesis, we propose a flexible framework combining pseudo-synthetic renderings from 3D city-wide meshes (*e.g.* Google Earth) with real, ground-level images (*e.g.* MegaDepth [29]). The pseudo-synthetic data is captured at varying altitudes, while the real, crowd-sourced images help improve visual fidelity especially for ground-level images where mesh-based renderings lack detail. The pipeline generates pseudo-synthetic images from different altitudes, co-registers them with real images, and aligns ground-level images with aerial data for 3D reconstruction. This hybrid dataset of real and pseudo-synthetic images provides geometric supervision that helps improve performance on downstream tasks such as ground-aerial camera registration and novel view synthesis, particularly in ground-aerial settings.

regression [62, 63, 72], etc. Datasets like BlendedMVS [69] generate training images by blending rendered images from textured 3D mesh with input real images. However, scenes in these datasets are *solely* captured from drones or ground, and not both simultaneously. This is because capturing data with drastic viewpoint changes is challenging and requires specialized sensors or manual effort. Instead, our pseudo-synthetic data, paired with real ground images, anchors scenes with simultaneous aerial and ground views.

Datasets with 3D city meshes [12, 24] have also been proposed. In contrast, our dataset contains both real and pseudo-synthetic data of the same landmarks in a unified coordinate system. While purely synthetic data [14, 28, 42, 52, 74] have been used for various 3D learning tasks, these datasets are not anchored to real scenes, making it difficult to bridge the sim-to-real gap. Methods using 3D meshes with image queries have been explored for other tasks like visual place recognition [4, 61] and localization [40, 41].

Learning for Multiview 3D. Driven by large-scale supervision, learning-based methods have resulted in more robust feature matching [10, 17–19, 33, 46, 58], pose estimation [62, 64, 71, 72] and direct regression of 3D pointmaps in a unified coordinate frame [26, 64]. Likewise, generating novel views from an input view conditioned on camera poses [34, 45, 56, 77] is enabled by large datasets like MegaScenes [60]. However, these methods struggle with extreme ground-aerial viewpoints due to limited supervision data. We show that our hybrid real and pseudo-synthetic data significantly improves performance of such learning-based methods on real mixed-altitude imagery.

Aerial-ground Registration. Prior work has addressed aerial-ground registration using specialized classical and learning-based methods [27, 31, 53, 78]. Our work is orthogonal, as we propose a hybrid dataset to potentially improve these methods among other tasks. In the same vein, very few co-registered ground-aerial datasets exist. MAVREC [16] and University1652 [75] provides no depth or pose information, and GrAco [79] has only monochrome imagery and is small-scale. We also distinguish our work from satellite-ground localization [20, 48, 54, 55, 66] as satellite viewpoint is orthographic and distant while viewpoints in our data are closer to a drone-view sharing common (albeit small) field-of-view with the ground-view.

3. Generating Aerial-Ground 3D Data

To address the scarcity of aerial-ground 3D data, we introduce an approach that combines renderings from 3D city-wide mesh models with real crowd-sourced images. The 3D meshes generate pseudo-synthetic renderings across varied altitudes and orientations, particularly aerial viewpoints, while real images complement this with ground-level cap-

tures, where pseudo-synthetic renderings often lack visual fidelity. An overview of our framework is in Figure 2.

3.1. Pseudo-synthetic data generation

We chose Google Earth as our primary data source for its quality and landmark coverage, allowing us to render images from any viewpoint. Our framework, however, is compatible with any geo-referenced 3D textured meshes [4, 9, 41]. These images are termed *pseudo-synthetic* as they are renderings of 3D meshes textured with real photos.

Automatically generating query viewpoints. Our goal is to render images that have sufficient visual overlap, both with each other and with the real images from MegaDepth [29]. We start with scenes from MegaDepth which contains SfM reconstructions for 196 landmarks, each with thousands of internet-sourced images. While the EXIF GPS tags in these images are not precise enough for accurate co-registration or alignment, we use them to roughly sample Google Earth's rendering viewpoints, ensuring they correspond to the same building or landmark. This is achieved by geo-referencing the local 3D reconstruction into the global ECEF frame using a similarity transform computed from the noisy GPS data. To ensure co-visibility among rendered images, we sample 200 points from the geo-referenced point cloud to serve as *look-at* targets for generating query viewpoints. This ensures that our rendered pseudo-synthetic images maintain sufficient visual overlap with each other and also with MegaDepth's real images.

Generating pseudo-synthetic 3D reconstruction. While users can specify and render images from any location and orientation, Google Earth unfortunately does not provide direct access to the underlying 3D mesh. Therefore, to recover scene geometry from the rendered pseudo-synthetic images (for which camera extrinsics and intrinsics are known), we extract keypoints [13] and match features [33] between the pseudo-synthetic rendered images to triangulate the 3D point cloud. With this process, we generated data from 137 sites or landmarks, each with 600 images taken from varying elevations, ranging from 1 meter to 350 meters, resulting in a total of 82,220 pseudo-synthetic images. We refer to this as *pseudo-synthetic reconstruction*.

3.2. Co-registering real crowd-sourced images

While pseudo-synthetic images from 3D textured meshes can directly improve geometry prediction tasks – as demonstrated in Section 5 – we make two key observations that help further improve the downstream performance. First, since these meshes are typically textured from aerial images, they tend to appear realistic from elevated, drone-like viewpoints but often lack visual fidelity at ground level, where low-viewing angles introduce artifacts on building facades [41]. A domain gap also exists between ground-level pseudo-synthetic and real images, such as the absence

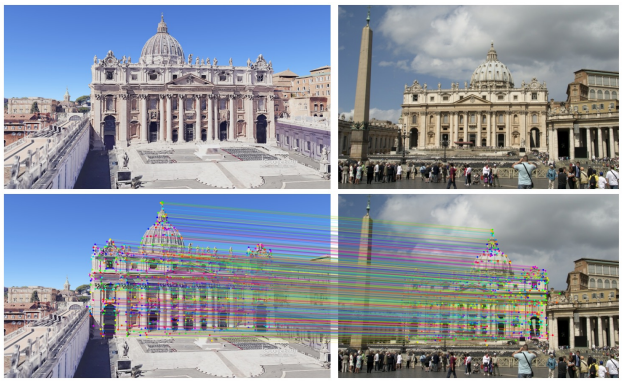

Figure 3. **Feature matching between real and pseudo-synthetic images.** The pseudo-synthetic rendering has a noticeable domain gap compared to the real MegaDepth image (e.g., no transients, simplistic lighting) but still enables reliable feature matching [46] to register real images into the pseudo-synthetic reconstruction.

of transients and simplistic lighting model, which may limit generalization to real-world data. Second, despite these limitations, mesh-based visual localization methods [41] show that real images can still be registered accurately to pseudo-synthetic reconstructions with state-of-the-art feature matching [33, 46, 58] as shown in Figure 3. To combine the benefits of both real and pseudo-synthetic images, we thus register real, ground-level images into the pseudo-synthetic reconstructions, resulting in aligned real (ground) MegaDepth and pseudo-synthetic (aerial) images.

Specifically, we follow standard visual localization pipeline [47] by first retrieving top-k most similar pseudo-synthetic images for each real MegaDepth query image [2]. 2D correspondences [13, 33] are lifted to 2D-3D matches using the pseudo-synthetic 3D points, and each query image's 6-DoF pose is estimated with a RANSAC-based PnP solver [25]. We refine the alignment by optimizing the localized MegaDepth images while keeping pseudo-synthetic cameras fixed. Using COLMAP's MVS [51], we generate semi-dense depth maps for supervision. In total, we register 49,937 MegaDepth images with 82,200 pseudo-synthetic images across 137 scenes, forming *AerialMegaDepth* – a hybrid dataset of 132,137 images with diverse viewpoints and lighting variations, as shown in Figure 4.

4. Learning Aerial-Ground 3D Reconstruction

We explore our data's impact on supervised learning for multi-view 3D reconstruction and novel view synthesis.

Selection of image pairs as supervision data. Our objective is to select image pairs that offer adequate overlap for effective supervision, particularly for ground-to-aerial pairs with altitude differences, ensuring the overlap is neither too high (making the task too easy) nor too low (making it too difficult). To achieve this, we compute an $N \times N$ covisibility matrix $\mathcal{C}$ for each scene, where each element

Figure 4. AerialMegaDepth data (top: MegaDepth, bottom: Google Earth) features **diverse viewpoints & lighting conditions**.

$\mathcal{C}[i,j] \in [0,1]$ represents the percentage of points in image i that are visible in image j. For ground-to-aerial settings, we prioritize pairs with significant viewpoint differences, meaning the covisibility is asymmetrical: we select pairs (i,j) such that $\mathcal{C}[i,j]$ is small and $\mathcal{C}[j,i]$ is large, or vice-versa. We quantify this difference with a score $s = \frac{\mathrm{AM}}{\mathrm{HM}}$, where AM is the arithmetic mean and HM is the harmonic mean of $\mathcal{C}[i,j]$ and $\mathcal{C}[j,i]$. A high score indicates pairs with a large viewpoint difference, ideal for challenging cross-view tasks, aligning with our goal of prioritizing ground-to-aerial pairs. Using this approach, we generate a varying-altitude dataset of 1.5M image pairs, each comes with camera intrinsics, camera poses, and depthmaps to be used as supervision for learning geometric tasks.

Multi-view Pose and Geometry Estimation. We consider the problem of estimating intrinsic, extrinsic camera parameters, and 3D scene from a set of N unconstrained images. Following the architecture of DUSt3R [64], we regress pointmaps for image pairs, where each pixel is associated with a 3D point in the coordinate system of the first frame. This allows us to compute the per-camera focal length (under a pinhole model with centered principal point) and the 6-DoF relative pose $\mathbf{T} = [\mathbf{R}|\mathbf{t}]$ for each image pair using PnP [25]. For $N > 2$ images, DUSt3R's global alignment (GA) step combines pointmaps across all images, optimizing a dense pairwise graph in 3D to align pointmaps within a global coordinate frame. We initialize with a DUSt3R checkpoint trained on millions of image pairs from eight datasets [11, 29, 37, 43, 49, 59, 69, 70] and fine-tune on our data, resulting in substantial improvements for ground-aerial camera registration. We also observe similar improvements by fine-tuning MASt3R [26] and using it as a front-end to provide 2D-2D correspondences, which are then fed into COLMAP for bundle adjustment (similar to the approach in MASt3R-SfM [15]).

Novel View Synthesis. For single-image novel view synthesis (NVS), our goal is to synthesize a plausible target ground view from a reference aerial image. Tung et al. [60] fine-tuned ZeroNVS [45] on MegaScenes with over 2M image pairs from 32K scenes, achieving significant improvements on scene-level view synthesis compared to object-centric settings [1, 43]. By further fine-tuning ZeroNVS on our dataset with varying altitudes, we achieve significant improvements in aerial-to-ground view synthesis.

5. Experiments

5.1. Multiview Pose Estimation and Reconstruction

Datasets. We fine-tune the baseline models [26, 64] on our data using pairwise 3D pointmaps as supervision. For evaluation, we focus on ground-aerial settings and include data from ULTRRA Challenge [67], which consists of images captured by ground-level cameras and drones, all calibrated using SfM constrained by RTK-corrected GPS coordinates for cm-level accuracy. Additionally, we use data from ACC-NVS1 [57] which captures various urban sites with ground-truth poses obtained via GNSS/IMU systems corrected by a stationary RTK base station. Overall, the evaluation data includes six sites with over 5,000 calibrated ground-aerial images.

Evaluation Metrics. Following [23, 62], we evaluate camera pose using *Relative Rotation Accuracy* (RRA) and *Relative Translation Accuracy* (RTA). RRA measures the angular difference between the predicted and ground-truth relative rotations, and RTA calculates the angular difference between the predicted and ground-truth translation vectors. We report $\mathrm{RTA@}\tau/\mathrm{RRA@}\tau$, i.e., the percentage of camera pairs with RTA/RRA below a threshold τ. We also evaluate reconstruction accuracy by aligning the predicted pointmap from DUSt3R with ground-truth from MVS [51] using a RANSAC-based optimal similarity transform. We report δ @ [0.5m, 1m, 2m], representing the percentage of points with errors within 0.5m, 1m, and 2m, respectively.

Two-view pose and geometry estimation. We evaluate the impact of our data on ground-aerial registration in the two-view case (one aerial and one ground image). For DUSt3R [64], relative pose is computed from the predicted 2D-3D matches using PnP [25]. We also present baselines of 2D correspondence matching with SuperPoint [13] + SuperGlue [46] (SP+SG), semi-dense matching with LoFTR [58], and MASt3R [26], where for MASt3R, 2D correspondences are extracted via reciprocal nearest neighbor from its dense local feature maps. For these methods, we compute relative pose assuming known ground-truth intrinsics, using essential matrix estimated from 2D matches.

Table 1 shows that all baseline methods, including SP+SG, LoFTR, DUSt3R and MASt3R, struggle with ground-aerial pairs. For example, baseline DUSt3R only recovers 5.20% of the total number of image pairs with good accuracy ($\mathrm{RRA@5°}$). While fine-tuning on synthetic data like MatrixCity [28] notably improves baseline methods, fine-tuning on pseudo-synthetic renderings (with more realistic textures, containing varying-altitude mesh renderings from Google Earth, denoted as DUSt3R/MASt3R +

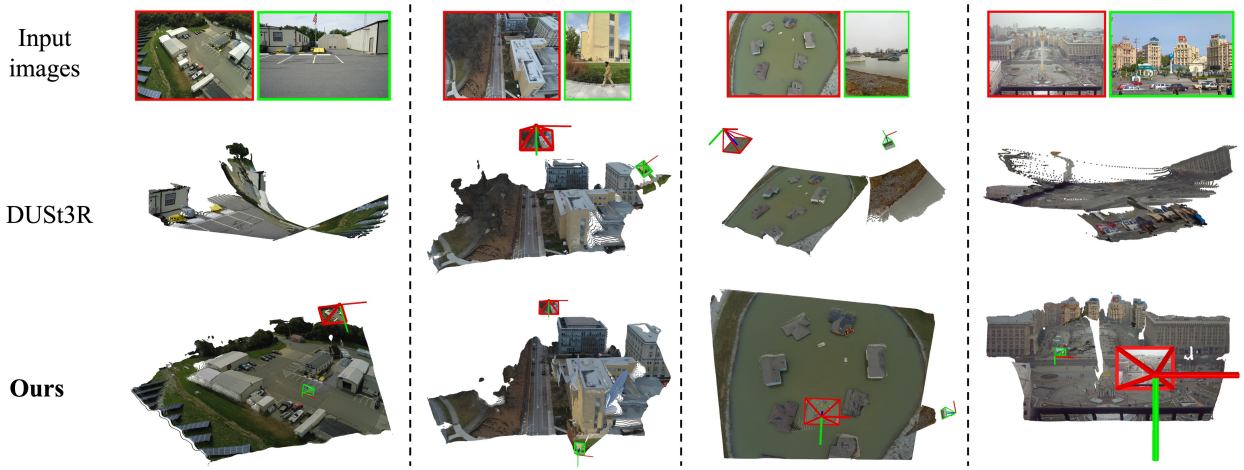

Figure 5. **Zero-shot ground-aerial camera and geometry prediction results.** Given two input images, one aerial and one ground, we compare the performance of the baseline DUSt3R [64] with the model fine-tuned on our varying-altitude data. The results demonstrate significant improvements over the baseline in **unseen, challenging** ground-aerial scenarios, showing the effectiveness of fine-tuning DUSt3R [64] with our data. Additionally, the last column presents qualitative results on a challenging ground-aerial pair from the WxBS [38] dataset, which involves significant viewpoint change.

Method	*Camera Rotation Accuracy*			*Camera Translation Accuracy*			*3D Pointmap Accuracy*		
	RRA@5°	RRA@10°	RRA@15°	RTA@5°	RTA@10°	RTA@15°	δ@0.5m	δ@1m	δ@2m
LoFTR [58]	0.92	1.83	2.45	0.92	1.53	2.14	-	-	-
SP+SG [46]	8.56	10.09	12.23	7.65	9.79	11.31	-	-	-
MASt3R [26] (released)	3.36	3.36	4.59	2.45	3.36	4.28	-	-	-
MASt3R + MatrixCity	19.78	30.88	38.17	10.67	25.72	29.43	-	-	-
MASt3R + PSynth (Ours)	26.49	43.71	47.62	25.25	40.32	49.34	-	-	-
MASt3R + Hybrid (Ours)	**49.54**	**66.36**	**72.48**	**42.51**	**63.30**	**69.11**	-	-	-
DUSt3R [64] (released)	5.20	7.95	9.48	2.75	5.81	9.17	29.02	42.16	43.79
DUSt3R + MatrixCity	17.85	37.28	42.80	11.33	25.24	33.24	31.43	47.13	57.02
DUSt3R + PSynth (Ours)	31.28	47.63	51.61	28.78	45.66	51.47	32.77	53.42	61.45
DUSt3R + Hybrid (Ours)	**55.96**	**71.25**	**76.15**	**46.48**	**68.20**	**72.78**	**38.24**	**62.33**	**74.52**

Table 1. **Finetuning with our data significantly improves pairwise camera pose estimation in the *ground-aerial* setting.** Baselines, including learned 2D correspondence matching (SP+SG [13, 46], LoFTR [58], MASt3R [26]) and 3D pointmap-based regression (DUSt3R [64]), struggle in this setting. For instance, DUSt3R localizes fewer than 5% of pairs within 5°rotation error (RRA@5°). Finetuning on MatrixCity improves performance, but using pseudo-synthetic ground-aerial pairs (DUSt3R + PSynth) boosts accuracy to 31%, and adding real ground data (DUSt3R + Hybrid) further increases it to 55%. This also significantly improves 3D pointmap accuracy. The first half of the table shows methods that predict 2D matches, with ground-truth intrinsics used to compute the relative poses.

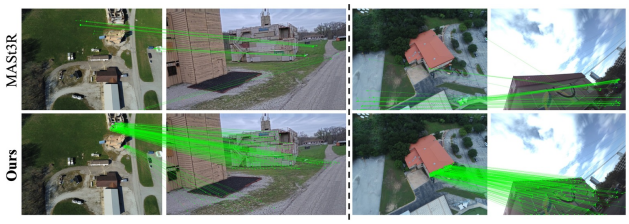

Figure 6. **Challenging ground-aerial feature matching.** Fine-tuned MASt3R [26] achieves accurate and robust feature matching across ground-aerial pairs with extreme viewpoint changes (correspondences extracted via reciprocal nearest neighbor from MASt3R's local feature maps). This highlights the effectiveness of our AerialMegaDepth data in improving matching performance.

PSynth) is more effective, increasing accuracy to 31.28% at RRA@5°. But the largest improvement comes from training on hybrid data (that aligns pseudo-synthetic mesh renderings to real-world images for pair construction), denoted as DUSt3R/MASt3R + Hybrid, bringing the performance to more than 55%. This demonstrates that our novel framework of hybrid real and pseudo-synthetic data significantly improves ground-aerial camera registration. We note that while our primary evaluation focuses on the ground-aerial setting, additional results in the Supplementary show that DUSt3R and MASt3R still perform well on similar-viewpoint pairs (e.g., ground-ground and aerial-aerial) even after fine-tuning with our varying-altitude data.

In addition to pose accuracy, we also observe substantial improvements in 3D geometry prediction, particularly in 3D pointmap accuracy from DUSt3R. As shown in Table 1, finetuning on our combined dataset (DUSt3R + Hybrid) improves the percentage of 3D points within a 1-meter error by 20% compared to the baseline DUSt3R. Since we align

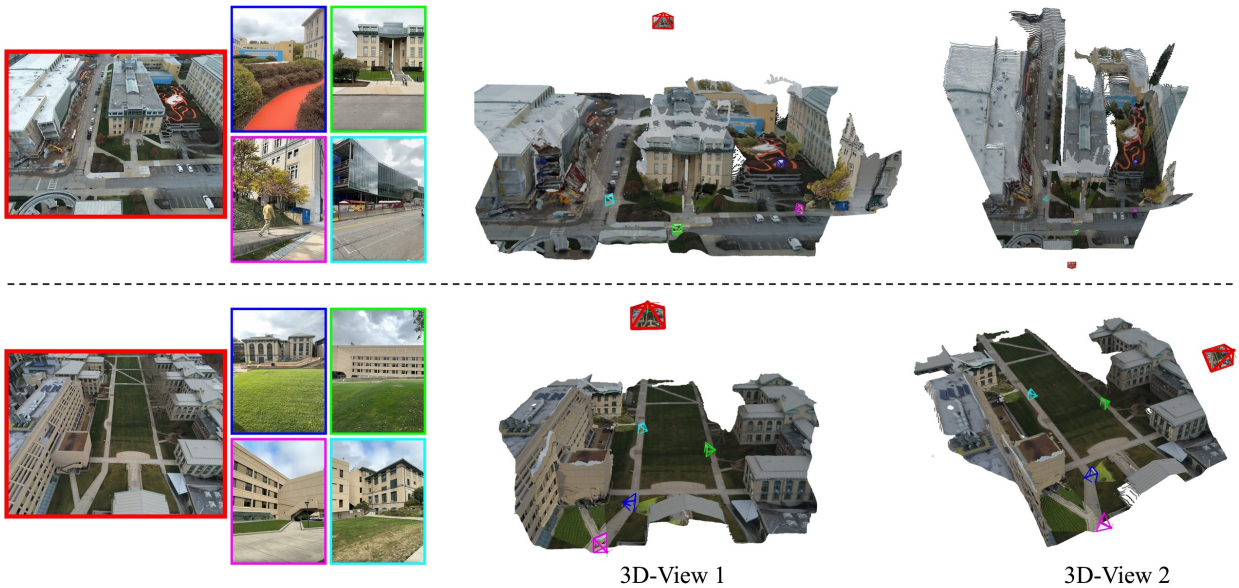

Figure 7. **3D reconstruction from one aerial and four ground images with *virtually no overlap*.** We use the global alignment process of DUSt3R [64] to merge pointmaps predictions. Despite the lack of overlap among the ground images, we find that incorporating a reference aerial image can effectively serve as a "map", significantly improving pose estimation accuracy when fine-tuned on our cross-view data.

# of ground images	2	3	4	5	6
no aerial image (i.e., ground only)					
DUSt3R-GA (released)	12.20	32.21	38.31	43.98	47.98
one aerial image					
DUSt3R-GA (released)	14.63	33.02	37.50	43.73	48.47
DUSt3R-GA + PSynth (Ours)	29.27	44.72	48.78	55.85	55.45
DUSt3R-GA + Hybrid (Ours)	**56.10**	**55.28**	**57.72**	**59.27**	**60.65**
MASt3R-SfM (released)	9.03	31.69	40.02	49.88	59.34
MASt3R-SfM + Psynth (Ours)	23.10	39.53	48.12	59.13	64.76
MASt3R-SfM + Hybrid (Ours)	**51.07**	**52.21**	**61.31**	**63.92**	**67.45**

# of ground images	2	3	4	5	6
no aerial image (i.e., ground only)					
DUSt3R-GA (released)	9.76	27.96	31.40	40.80	43.13
one aerial image					
DUSt3R-GA (released)	9.76	27.15	31.40	41.78	43.62
DUSt3R-GA + PSynth (Ours)	31.27	43.09	46.82	55.72	56.10
DUSt3R-GA + Hybrid (Ours)	**51.29**	**52.85**	**54.07**	**55.61**	**57.72**
MASt3R-SfM (released)	9.28	23.91	29.01	46.91	51.45
MASt3R-SfM + Psynth (Ours)	25.80	41.09	44.52	58.79	61.22
MASt3R-SfM + Hybrid (Ours)	**48.84**	**49.71**	**57.89**	**60.98**	**62.41**

Table 2. **Including a single aerial image with N ground images notably improves pose estimation of the ground images**, as shown in Ground Cameras Rotation Accuracy @ 15° (RRA@15°) (left) and Translation Accuracy @ 15° (RTA@15°) (right). Using DUSt3R's global optimization [64], Row 1 shows results for ground-only input images, while the rest includes an aerial image as input. Although pose estimation improves with more ground images, adding even one aerial reference image significantly boosts accuracy, especially when ground images have minimal overlap (e.g., $N \leq 3$) as this aerial view helps align the ground images within a shared coordinate frame.

predicted and ground-truth pointmaps using a RANSAC-based similarity transform, baseline DUSt3R model could still achieve reasonable geometry accuracy by producing good depth estimates for at least one of the views, *even if they struggle to register them together accurately* (as shown in pose accuracy). By finetuning on our data, we see substantial improvements in both 3D pointmap accuracy and ground-aerial pose registration, highlighting the impact on both pose and geometry prediction. We show qualitative results in Fig. 5 and Fig. 6 and encourage readers to check the website for additional results. We emphasize that this is zero-shot performance on unseen data, as there is no overlap between our training and evaluation scenes.

Co-registering ground images with aerial context. We evaluate multi-view ground-aerial camera registration, where one aerial image is matched with N ground images. To perform multiview pose estimation, we explore two approaches: 1) the global alignment (GA) process of DUSt3R [64] to merge all predicted pairwise pointmaps in the same coordinate frame, and 2) the MASt3R-SfM [15] approach, which uses MASt3R as a front-end to extract 2D correspondences followed by COLMAP [50] bundle adjustment. In Table 2, we report accuracy for ground images only. Perhaps unsurprisingly, when ground images are sparse ($N \leq 3$), they often lack visual overlap, making pose estimation particularly challenging. In this case, we observe a significant improvement in performance (from 14.63% to 56.10% for RRA@15° with $N = 2$) when using the model finetuned with our data. The key reason for this improvement is, even with almost no overlap among ground images, a single aerial image can serve as an "overhead map", helping to "stitch" or align the ground images into a common

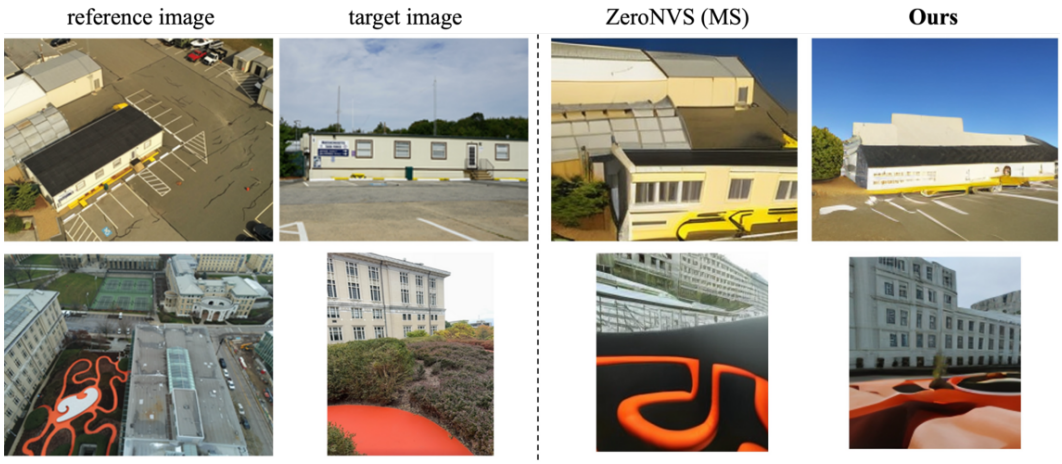

Figure 8. **Results of extreme viewpoint change in novel-view synthesis** with ZeroNVS [45] finetuned on MegaScenes [60] (ZeroNVS MS) & additionally finetuned on our data. Though by no means perfect, note the big improvement in visual quality and viewpoint accuracy.

NVS	DreamSim ↓	LPIPS ↓	PSNR ↑	SSIM ↑
pseudo-synth. images				
ZeroNVS (MS)	0.448	0.413	10.847	0.416
ZeroNVS (Ours)	**0.377**	**0.359**	**12.381**	**0.484**
real images				
ZeroNVS (MS)	0.550	0.639	7.478	0.183
ZeroNVS (Ours)	**0.442**	**0.580**	**8.220**	**0.218**

Table 3. **Quantitative results for aerial-ground novel-view synthesis** comparing ZeroNVS model finetuned on MegaScenes (MS) and our data (Ours), with finetuning improves all metrics.

frame. An example is shown in Figure 7 demonstrating the effectiveness of our approach in this scenario.

5.2. Novel View Synthesis

Datasets. We focus on the challenging aerial-to-ground synthesis task. We combine our dataset with MegaScenes [60], using a 3:1 ratio during finetuning to help prevent overfitting. For evaluation, we use both real-world aerial-ground pairs as well as pseudo-synthetic data from Google Earth that includes a wide range of images captured at varying altitudes, allowing us to assess the model's ability to synthesize views across diverse ground-aerial setups.

Finetuning details. We follow ZeroNVS [45] for novel view synthesis, which takes the extrinsic matrix and field-of-view as inputs to generate novel views at the target pose. The translation vector is scaled based on the 20th depth quantile of the reference image (during evaluation we used MVS depth). Starting from ZeroNVS trained on MegaScenes, we fine-tune the model on our dataset, significantly improving novel views in ground-aerial contexts.

Evaluation Metrics. We evaluate view synthesis quality using standard image reconstruction metrics, including LPIPS [73], PSNR, and SSIM [65]. Additionally, we also include the DreamSim [21] score, which aligns more closely with human perceptual judgments.

Results and Discussions. Table 3 presents significant quantitative improvements for single image aerial-to-ground novel-view synthesis. From the qualitative results in Figure 8, we see that ZeroNVS (Ours), produces realistic and accurate images that follow the desired poses. In contrast, ZeroNVS (MS), which was finetuned solely on MegaScenes, struggles with such views, highlighting once again the effectiveness of incorporating ground-aerial data into the training process. We emphasize that this is still a very challenging task, as the viewpoint difference between the reference and target pose is large. The network must learn to retain the underlying scene structure while generating plausible images for unseen parts of the scene and/or demonstrate correct occlusions. Our results show that the model somewhat successfully addresses this challenge, but much research remains to be done in this task.

6. Conclusion

Despite notable advances in learning-based 3D reconstruction, large-area reconstruction from a sparse mix of drone and ground imagery remains a challenge. As shown over the last decade, adding significant data where little existed before improves the performance of supervised-learning-based networks. The key innovation in our work comes from understanding how geospatial platforms and crowd-sourced imagery can be combined to provide a potentially unlimited amount of data for training large aerial-ground 3D models. Carefully finetuning existing 3D models with our data showed nearly 15× improvement in camera estimation and registration, which is at the heart of the large-scale reconstruction problem. We hope our hybrid data framework will help spur further research in the area.

In the future, aerial drone views could serve as a bridge between ground and satellite views, where abundant data is widely available. Combining them all could bring us closer to the ambitious goal of planet-scale 3D reconstruction.

Acknowledgements: This work was supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. 5

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016. 4

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 2

[4] Gabriele Berton, Lorenz Junglas, Riccardo Zaccone, Thomas Pollok, Barbara Caputo, and Carlo Masone. Meshvpr: Citywide visual place recognition using 3d meshes. In *ECCV*, 2024. 3, 4

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 2

[6] Bing. Bing Streetside. https://www.bing.com/maps/. 2

[7] 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

[8] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *ICCV*, 2023. 2

[9] Cesium. Cesium. https://cesium.com/. 2, 4

[10] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. *ECCV*, 2022. 3

[11] Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *NeurIPS Datasets and Benchmarks*, 2021. 5

[12] Jean-Emmanuel Deschaud, David Duque, Jean Pierre Richa, Santiago Velasco-Forero, Beatriz Marcotegui, and François Goulette. Paris-carla-3d: A real and synthetic outdoor point cloud dataset for challenging tasks in 3d mapping. *Remote Sensing*, 2021. 2, 3

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised Interest Point Detection and Description. In *CVPR*, 2018. 2, 4, 5, 6

[14] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017. 3

[15] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. In *3DV*, 2025. 5, 7

[16] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception? In *CVPR*, 2024. 3

[17] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023. 3

[18] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Dedode: Detect, don't describe—describe, don't detect for local feature matching. In *3DV*, 2024.

[19] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *CVPR*, 2024. 2, 3

[20] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *CVPR*, 2023. 3

[21] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *NeurIPS*, 2024. 8

[22] Google. Google Street View. https://www.google.com/streetview/. 2

[23] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2021. 2, 5

[24] Binyu Lei, Rudi Stouffs, and Filip Biljecki. Assessing and benchmarking 3d city models. *International Journal of Geographical Information Science*, 2023. 2, 3

[25] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *IJCV*, 2009. 4, 5

[26] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2, 3, 5, 6

[27] Hongjie Li, Aonan Liu, Xiao Xie, Han Guo, Hanjiang Xiong, and Xianwei Zheng. Learning dense consistent features for aerial-to-ground structure-from-motion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023. 3

[28] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *ICCV*, 2023. 3, 5

[29] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 2, 3, 4, 5

[30] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *3DV*, 2024. 2

[31] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *CVPR*, 2015. 3

[32] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021. 2

[33] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, 2023. 2, 3, 4

[34] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2, 3

[35] David G. Lowe. Distinctive Image Features from Scale-invariant Keypoints. *IJCV*, 2004. 2

[36] Mapillary. Mapillary Maps. https://www.mapillary.com/. 2

[37] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 5

[38] Dmytro Mishkin, Jiri Matas, Michal Perdoch, and Karel Lenc. Wxbs: Wide baseline stereo generalizations. In *BMVC*, 2015. 6

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

[40] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *ECCV*, 2022. 3

[41] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Visual localization using imperfect 3d models from the internet. In *CVPR*, 2023. 3, 4

[42] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *CVPR*, 2023. 3

[43] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 5

[44] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: an efficient alternative to SIFT or SURF. In *ICCV*, 2011. 2

[45] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. *CVPR*, 2023. 1, 2, 3, 5, 8

[46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 3, 4, 5, 6

[47] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 4

[48] Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. Snap: Self-supervised neural maps for visual positioning and semantic understanding. In *NeurIPS*, 2023. 3

[49] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 5

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 7

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 4, 5

[52] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, 2018. 3

[53] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Accurate geo-registration by ground-to-aerial image matching. In *3DV*, 2014. 3

[54] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *CVPR*, 2022. 3

[55] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *NeurIPS*, 2019. 3

[56] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[57] Thomas Sugg, Kyle O'Brien, Lekh Poudel, Alex Dumouchelle, Michelle Jou, Marc Bosch, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Accenture-nvs1: A novel view synthesis dataset. *arXiv:2503.18711*, 2025. 5

[58] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 2, 3, 4, 5, 6

[59] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 5

[60] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 1, 2, 3, 5, 8

[61] Andrea Vallone, Frederik Warburg, Hans Hansen, Søren Hauberg, and Javier Civera. Danish airs and grounds: A dataset for aerial-to-street-level place recognition and localization. *Robotics and Automation Letters*, 2022. 3

[62] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 2, 3, 5

[63] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, 2024. 2, 3

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 2, 3, 5, 6, 7

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 8

[66] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *ICCV*, 2015. 3

[67] Workshop. Ultrra: Unconstrained large-scale three-dimensional reconstruction and rendering across altitudes. In *WACV 2025*. 5

[68] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[69] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 3, 5

[70] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 5

[71] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 3

[72] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *ICLR*, 2024. 2, 3

[73] 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. 8

[74] Saining Zhang, Baijun Ye, Xiaoxue Chen, Yuantao Chen, Zongzheng Zhang, Cheng Peng, Yongliang Shi, and Hao Zhao. Drone-assisted road gaussian splatting with cross-view uncertainty. In *BMVC*, 2024. 3

[75] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. *ACM Multimedia*, 2020. 3

[76] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, 2021. 2

[77] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3

[78] Qing Zhu, Zhendong Wang, Han Hu, Linfu Xie, Xuming Ge, and Yeting Zhang. Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3d reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020. 3

[79] Yilin Zhu, Yang Kong, Yingrui Jie, Shiyou Xu, and Hui Cheng. Graco: A multimodal dataset for ground and aerial cooperative localization and mapping. *Robotics and Automation Letters*, 2023. 3