ULTRA-360: UNCONSTRAINED DATASET FOR LARGE-SCALE TEMPORAL 3D RECONSTRUCTION ACROSS ALTITUDES AND OMNIDIRECTIONAL VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Significant progress has been made in photo-realistic scene reconstruction over recent years. Various disparate efforts have enabled capabilities such as multi-appearance or large-scale reconstruction from images acquired by consumer-grade cameras. How far away are we from digitally replicating the real world in 4D? So far, there appears to be a lack of well-designed dataset that can evaluate the holistic progress on large-scale scene reconstruction. We introduce a collection of imagery on a campus, acquired at different seasons, times of day, from multiple elevations, views, and at scale. To estimate many camera poses over such a large area and across elevations, we apply a semi-automated calibration pipeline to eliminate visual ambiguities and avoid excessive matching, then visually verify all calibration results to ensure accuracy. Finally, we benchmark various algorithms for automatic calibration and dense reconstruction on our dataset, named ULTRA-360, and demonstrate numerous potential areas to improve upon, e.g., balancing sensitivity and specificity in feature matching, densification and floaters in dense reconstruction, multi-appearance overfitting, etc. We believe ULTRA-360 can serve as the benchmark that reflect realistic challenges in an end-to-end scene-reconstruction pipeline.

Figure 1: Sample images of one building, Building #10, in our imagery collected over multiple seasons, elevations, and multiple camera types to enable fully immersive 3D/4D reconstruction.

1 INTRODUCTION

Immersive digitalization of the 3D world is of great interests for Computer Vision and Graphics researchers, with many real-world applications in Robotics, AR/VR, Autonomous Driving, Urban Planning, etc. Tremendous progress has been made in recent years with neural rendering innovations. Methods such as Neural Radiance Field (NeRF) (Mildenhall et al., 2020) and 3D Gaussian

Splatting (Kerbl et al., 2023) have improved quality in photorealistic scene reconstruction and novel-view synthesis. Various follow-up works have further extended this capability at larger scale (Turki et al., 2022; Liu et al., 2024a; Tancik et al., 2022b), to model sites across time (Martin-Brualla et al., 2021; Chen et al., 2022; Yang et al., 2023; Kulhanek et al., 2024b; Xu et al., 2024c; Zhang et al., 2024), and in various challenging scenarios, such as sparsity, incorrect exposure, or on degraded images (Zhu et al., 2023; Tang et al., 2025; Peng & Chellappa, 2023; Gao et al., 2024; Wu et al., 2025). With rapid progress and a vast amount of visual data online, we are closer than ever to achieving immersive 3D, perhaps even 4D, digitalization of the world. However, the current advances in neural reconstruction are often measured with disparate benchmarks and in specific aspects; collections of data from the internet (Snavely et al., 2006; Wallingford et al., 2024) are also difficult to be used for accurate evaluations, given the myriad of uncontrollable variables within the collection.

Pose and the Structure-from-Motion (SfM) point cloud from camera calibration are essential to dense reconstruction. Inverse rendering work typically assumes *known camera poses*, but how realistically can we assume accurate camera calibration, particularly for large scale scenes? Methods that address multi-view inconsistencies (Martin-Brualla et al., 2021; Chen et al., 2022; Yang et al., 2023; Kulhanek et al., 2024b; Xu et al., 2024c; Zhang et al., 2024) work well on small scale scenes with *dense* camera coverage. Are these methods adaptable to sparser coverage or in large scale scenes? While open source efforts (Tancik et al., 2023; Yu et al., 2022) have attempted to accommodate different methods and datasets, they have largely stalled as software complexity grows over time. How realistically can we reconstruct city-scale scenes with only 2D images, in an end-to-end manner? We summarize two limitations in the current benchmark datasets for calibration, dense reconstruction and Novel View Synthesis (NVS):

Lack of Scale in Camera Coverage. Current datasets are typically limited in two areas: the perspective camera format and the limited camera coverage. Perspective cameras are ubiquitous and easy to use; however, their limited Field-of-Views (FoV) lead to partial observation of the scene. As such, reconstructions are only viewable in one direction and are undesirable for immersive exploration. The distribution of cameras are also focused on one aspect of the scene, e.g. either on the ground (Tancik et al., 2022a; Meuleman et al., 2023) or in the air (Turki et al., 2022; Crandall et al., 2011). While aerial observations can recover large scale structures, ground observations contain much richer details. Additionally, the limited camera coverage makes NVS evaluation overly reliant on test cameras that are close to training cameras, and does not reveal issues such as obvious floaters (Warburg et al., 2023) in unconstrained exploration of the 3D asset.

Lack of Scale in Realism and Time. Synthetic data from unlimited perspectives and FoV can be generated from virtual engines (Li et al., 2023; Xiangli et al., 2022; Mittal et al., 2023), but such data lacks realism. Real images are full of inconsistencies that cannot be fully simulated, e.g., lighting, seasonality, weather, etc. So far, datasets that demonstrate these realistic scenarios (Snavely et al., 2006; Sabour et al., 2023) are small in scale and difficult to evaluate against. For example, Phototourism (Snavely et al., 2006) comprises of internet images collected at unknown time. As a result, methods (Martin-Brualla et al., 2021; Xu et al., 2024b; Kulhanek et al., 2024a; Xu et al., 2024a) developed on these datasets requires access to *test-view* images during evaluation to optimize appearance information. Various temporal concepts such as seasonality and structural modifications are neither fully transient nor based on only appearance changes.

We propose a dataset for Unconstrained Large-scale Temporal 3D Reconstruction across Altitudes, named *ULTRA-360*. ULTRA-360 is collected at a campus and aims to reconstruct and visualize the entire campus in 4D, with *hundreds of videos* collected across the span of *two years*, where the video frames are calibrated with *manual inspection* and aligned to a *consistent coordinate system*. ULTRA-360 provides:

1. **Immersive Ground Acquisition**: Both *perspective* and *360 panorama* images on the ground level are collected and calibrated to facilitate immersive 3D reconstruction.
2. **Multi-Elevation Acquisition**: Both *ground* and *aerial* images from multiple elevations are collected and calibrated to ensure full coverage of the buildings.
3. **Multi-Seasonality Acquisition**: Images are acquired across multiple seasons in a two year period, capturing the gradual changes over time.
4. **Large-Scale Calibration**: Twenty academic buildings are collected across multiple elevations, camera models, years, and are calibrated together.

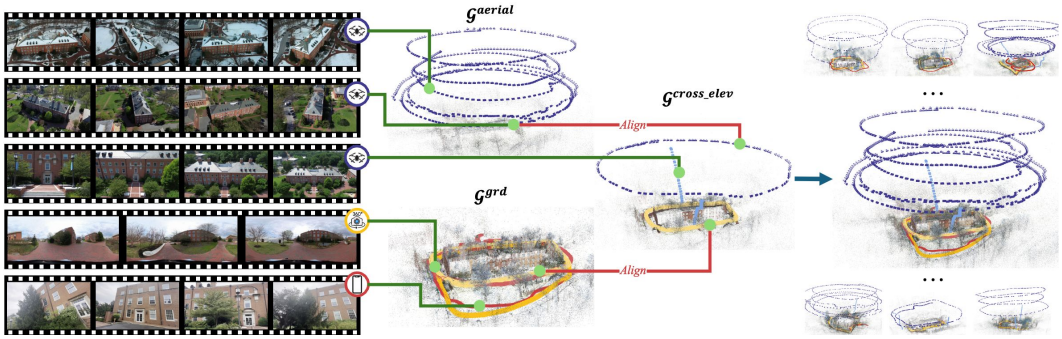


Figure 2: Visualization of the large-scale camera calibration process.

We perform detailed evaluations of current State-of-The-Art methods on ULTRA-360, both in feature matching and dense reconstruction. The results reveal encouraging process and challenges to be addressed. For feature matching, recent innovations allow us to find correspondences across large distances. Despite such progress, SoTA feature matching methods (Lowe, 2004; DeTone et al., 2018; Edstedt et al., 2024b; Sun et al., 2021; Leroy et al., 2024; Sarlin et al., 2020) lie between the spectrum of *insufficient true positives* between images with large baselines and *significant false positives* between images with visual ambiguities. Scene graph optimization techniques (Cai et al., 2023; Xiangli et al., 2024; Arandjelovic et al., 2018; Duisterhof et al., 2024) can ameliorate such a process, but still requires various manual intervention.

Dense reconstruction from multiple elevations suffers from difficulties in sufficient densification and severe sky floaters, despite improvements to Level-of-Details. Multi-appearance modeling is often entangled with view-direction bias if treated as a per-image optimization. We make several modifications, including neural sky modeling and time-based appearance modeling, to tackle these issues, and expect future research to improve immersive reconstruction based on ULTRA-360.

2 Related Work

Table 1: A comparison of existing multi-view datasets highlighting key properties, including scale, diversity of appearances, FoV, and viewpoint variation.

Dataset	# images	Scale	Real/Synthetic	Time	Camera Type	Elevation
Phototourism (Snavely et al., 2006)	150K	Scene	Real	Uncontrolled	Perspective	Ground
MegaScenes (Tung et al., 2024)	2M	Scene	Real	Uncontrolled	Perspective	Ground
BlendedMVS (Yao et al., 2020)	5K	Scene	Real+Synthetic	Single	Perspective	Ground
UrbanScene3D (Crandall et al., 2011)	128K	Scene	Real+Synthetic	Single	Perspective	Aerial
Quad 6K (Crandall et al., 2011)	5.1K	Scene	Real	Single	Perspective	Aerial
Mill 19 (Turki et al., 2022)	3.6K	Scene	Real	Single	Perspective	Aerial
OMMO (Lu et al., 2023)	14.7K	Scene	Real	Day/Night	Perspective	Aerial
Block-NeRF (Tancik et al., 2022b)	2.8M	City	Real	Day/Night	360	Ground
KITTI-360 (Liao et al., 2023)	300K	City	Real	Single	360	Ground
NuScenes (Caesar et al., 2020)	1.4M	City	Real	Day/Night/Rainy	360	Ground
MatrixCity (Li et al., 2023)	519K	City	Synthetic	Diff. Weather/Lighting	Perspective	Ground+Aerial
ULTRA-360	37.7 K	City	Real	Four Seasons, Day/Night	Perspective+360	Ground+Aerial

2.1 Multi-View Datasets for Dense Reconstruction

In scene reconstruction and NVS research, the widely used benchmark datasets often focus on single objects (Mildenhall et al., 2020; Knapitsch et al., 2017; Barron et al., 2022) or indoor scenes (Lin et al., 2018). These datasets are collected in controlled environments with accurate camera estimation. Various datasets (Snavely et al., 2006; Tung et al., 2024) construct outdoor unbounded architecture datasets with multi-view images from the internet. While these datasets include appearance diversity, the uncontrolled collection method leads to lack of multi-view imagery on a single consistent appearance. Consequently, algorithms (Martin-Brualla et al., 2021; Chen et al., 2022; Yang et al., 2023; Kulhanek et al., 2024b; Xu et al., 2024c; Zhang et al., 2024) tested on these datasets require access to test-view images during evaluation to account for unique appearance variation.

Large-scale datasets, such as Quad 6K (Crandall et al., 2011), UrbanScene3D (Crandall et al., 2011), Mill-19 (Turki et al., 2022), and OMMO (Lu et al., 2023), have been collected from an aerial platform. This limits the level of details in reconstructed models, if rendering or exploration from the

ground perspective is desired. Driving datasets like Block-NeRF (Tancik et al., 2022b), KITTI-360 (Liao et al., 2023), and NuScenes (Caesar et al., 2020) focus on street-level imagery, leading to many unobserved regions such as the roof of the buildings. So far, no dataset has been proposed for a large-scale collection of imagery spanning multiple elevations. MatrixCity (Li et al., 2023) contains both ground and aerial imagery, but is synthesized through game engines.

2.2 Camera Calibration and Dense Reconstruction Algorithms

Recovering dense 3D geometry from 2D images has a long history of research. Broadly speaking, camera calibration is first performed based on SfM (Schönberger & Frahm, 2016; Schönberger et al., 2016; Pan et al., 2024), which relies on reliable feature extractors (DeTone et al., 2018; Lowe, 1999; Edstedt et al., 2024a; 2025) and feature matchers (Sarlin et al., 2020; Lindenberger et al., 2023b; Sun et al., 2021; Edstedt et al., 2024b; Leroy et al., 2024) to find correspondences. SfM then performs triangulation to recover camera pose and sparse geometry. Such a calibration process can be computationally expensive or get stuck in incorrect solutions due to visual ambiguities; various scene graph optimization techniques (Arandjelovic et al., 2018; Berton et al., 2023; Cai et al., 2023; Xiangli et al., 2024) have been introduced to remove unnecessarily or ambiguous image pairs based on prior knowledge. Selecting the proper scene graph or feature matching algorithms are highly subjective and unpredictable; while datasets from the Image Matching Challenging (Bellavia et al., 2025) exist, they are not constructed to be also used for dense reconstruction evaluation.

Dense reconstruction, or photorealistic NVS, has progressed significantly with the introduction of NeRF (Mildenhall et al., 2020) and 3DGS (Kerbl et al., 2023). By optimizing an implicit or explicit radiance field on multi-view images through differential rendering, these methods can achieve photorealistic rednering quality. Follow-up works has improved upon NeRF (Turki et al., 2022; Tancik et al., 2022a; Mi & Xu, 2023; Reiser et al., 2023; Xiangli et al., 2022; Meuleman et al., 2023) and 3DGS (Lin et al., 2024; Liu et al., 2024b; Lu et al., 2024; Ren et al., 2024) in large scale reconstruction, e.g., by splitting the scene into multiple blocks for optimization, introducing Level-of-Detail rendering, multi-appearance modeling, etc. Evaluation is done on test cameras, typically in-between training cameras; however, no quantitative evaluation has been done on more free-formed and realistic novel views. As shown in Table 1, ULTRA-360 provides rich variations in both appearances, rendering FoVs, and cameras ranging from the ground to the sky, providing an unique opportunities to understand the effects of view-dependent effects, floaters, and details.

3 ULTRA-360

As shown in Figure 2, ULTRA-360 captures real-world, large-scale imagery with multi-appearance, multi-elevation, panorama coverage, and providing a comprehensive testing ground for evaluating modern scene reconstruction and NVS algorithms. This dataset contains over *30k calibrated images* on twenty academic halls within a campus, covering an area of approximately *140 acre* and a time period of *two years*. ULTRA-360 covers a variety of texture and material, e.g., grass, glass/windows, trees, rocks, etc., that are on the campus. In the following section, we describe the data collection process and the semi-automated calibration pipeline to construct ULTRA-360.

3.1 Large-scale Data Collection across Time and Elevation in 360 Degrees

Table 2: Summary of ULTRA-360, where multi-view sequences are collected at different time, appearances, elevations, and FoVs.

Device	# Videos	# Frames	Season	Appearance	FoV	Elevation
iPhone	19	7134	Summer, Fall	Sunny, Cloudy, Night	70°	0m
Insta360	31	23260	Spring, Winter	Sunny, Cloudy, Night	360°	0m
DJI Mini 3	81	7334	Spring, Winter	Sunny, Cloudy, Night	82°	60, 100, 120m

Constructing a dataset for large scale, immersive 3D reconstruction over time is laborious, time-consuming, and computationally intensive. While professional photogrammetry software and devices exist, they are not scalable and difficult to integrate with novel research. To enable collection at scale in coverage and time, we elect to use a variety of consumer-grade devices and develop our own processing pipeline. As shown in Table 2, for each of the twenty buildings, we systematically collect both aerial and ground-level imagery across four seasons with different lighting conditions.

For **ground** imagery, the data collection process involves walking around each building's perimeter with an iPhone or Insta360 camera to capture video sequences. We perform manual inspection on all extracted frames to remove low-quality images and ensure sufficient overlap. Particularly, panorama frames are split into four perspective images, each with a 120° FoV. These four frames together cover the horizontal 360° FoV around the camera. We discard the bottom face, which has a static human operator, and the top face, which is mostly sky. Any image that contains Personally Identifiable Information (PII), e.g., faces or vehicle license plates, are blurred through automated algorithms (Wu et al., 2019).

For **aerial** imagery, we operate DJI drones that follow a circular flight trajectory around the building. Drone flights are planned to ensure uniform and complete coverage. Multiple elevations are collected at 60, 100, and 120m. We also keep the ascending video sequences as the drone moves from ground level to approximately 60m above ground on two sides of each building. These ascending videos help improve calibration between ground and aerial imagery. From these videos, we sample individual frames, applying the same quality control measures as for ground-level data.

3.2 Semi-Automated Camera Calibration for Doppelganger Mitigation

After video acquisition and frame extraction, we build a semi-automated pipeline to obtain correct camera calibration for all images. Given the sheer size in the number of images and covered area, directly applying software, e.g., COLMAP (Schönberger & Frahm, 2016), is both infeasible and will lead to inaccurate results. As shown in Figure 2, we use a divide-and-conquer approach by 1. calibrating images within an elevation, 2. merging images across multiple elevations based on a manually verified cross-elevation set, and 3. merging images from different buildings into a single coordinate system.

Image Calibration within an Single Elevation. For camera calibration, a collection of images $\mathcal{I}$ are collected at different times. Based on these images, scene graphs $\mathcal{G}^{\text{grd}}$ and $\mathcal{G}^{\text{aerial}}$ can be constructed from the ground and aerial images. Scene graphs $\mathcal{G} = (\mathcal{I}, \mathcal{P})$ consist of $\mathcal{I}$ as nodes, and image pairs $\mathcal{P} = \{(\mathcal{I}_i, \mathcal{I}_j)\}$ as edges. Correspondences between $(\mathcal{I}_i, \mathcal{I}_j)$ are extracted if edge $\mathcal{P}$ exists; such correspondences are then used for triangulation in SfM. I.e., $\mathcal{G}$ determines the visibility of $\mathcal{I}$ to other images. Various implementations can be used to determine scene graph edges. Exhaustive scene graphs are generally more accurate, but can also lead to more false matches.

Visual ambiguity, often referred to as doppelgänger (Cai et al., 2023) matches, occur to cameras that are far apart due to their similar patterns. These doppelgängers are particularly common for ground image collection of *buildings*. Aerial images suffer less from visual ambiguities, as they have a more global view of the building. For $\mathcal{G}^{\text{aerial}}$, we simply use exhaustive matching. For $\mathcal{G}^{\text{grd}}$, we use a mixture of sequential and exhaustive scene graph constructions to avoid doppelgängers.

Specifically, we denote multi-appearance ground images as $\mathcal{I}_i^x$, where x denotes the video sequence and i denote the frame within the sequence. Image pairs $\mathcal{P} = \{\mathcal{P}^x_{\text{within}}\} \cup \{\mathcal{P}^{x,y}_{\text{between}}\}$ can be fully separated into pairs that are within sequence x and between any two sequences $\{x, y\}$. For $\mathcal{P}^x_{\text{within}}$, we use sequential matching, i.e. $\mathcal{P}^x_{\text{within}} = \{(\mathcal{I}_i^x, \mathcal{I}_j^x) || i - j| \leq 10\}$, which prevents far-away frames to match. Such a spatial constraint is harder to determine for $\mathcal{P}^{x,y}_{\text{between}}$, as different sequences may not follow the same path or pace. To this end, we manually bucket frames into $\mathcal{S}^x_{\text{front}}$ and $\mathcal{S}^x_{\text{back}}$, which denote frames that are looking at the *front* or *backside* of the building. $\mathcal{P}^{x,y}_{\text{between}}$ can then be define as:

$$\mathcal{P}^{x,y}_{\text{between}} = \{(\mathcal{I}_i^x, \mathcal{I}_j^y) | i \in \mathcal{S}^x_{\text{front}}, j \in \mathcal{S}^y_{\text{front}}\} \cup \{(\mathcal{I}_i^x, \mathcal{I}_j^y) | i \in \mathcal{S}^x_{\text{back}}, j \in \mathcal{S}^y_{\text{back}}\}. \tag{1}$$

We find this setup effectively eliminates cross-sequence doppelgangers, as visual ambiguity within the same side of the building can be controlled by spatial constraints of individual sequences. For panorama images, which are split into four perspective frames, $\mathcal{P}^{x,y}_{\text{between}}$ against iPhone frames only involve the building-facing side of the panorama image.

Cross-Elevation Calibration. To connect calibrations from different elevations, we perform an additional calibration on a cross-elevation set. Specifically, we calibrate a panorama ground sequence with an aerial sequence. Registering cameras across a large baseline is challenging, due to a lack of sufficient correspondences. To assist cross-elevation calibration, we record two transitional drone sequences from ground to air for each building. Similar to ground images, transitional drone images can experience visual ambiguities at ground level. The two sequences are distributed on the front

and backside of the building. We manually define the scene graph $\mathcal{G}^{\text{cross\_elev}}$, i.e.,

$$\mathcal{P}^{\text{cross\_elev}} = \{\mathcal{P}^{\text{grd}}_{\text{grd}}\} \cup \{\mathcal{P}^{\text{grd}}_{\text{trans}}\} \cup \{\mathcal{P}^{\text{grd}}_{\text{aerial}}\} \cup \{\mathcal{P}^{\text{trans}}_{\text{trans}}\} \cup \{\mathcal{P}^{\text{trans}}_{\text{aerial}}\} \cup \{\mathcal{P}^{\text{aerial}}_{\text{aerial}}\}, \tag{2}$$

where $\mathcal{P}^{x}_{y}$ denotes image pairs between two elevations (note that $\mathcal{P}^{x}_{y} \equiv \mathcal{P}^{y}_{x}$). For ground images, we apply sequential matching similar to the ground-only scenario previously, i.e., $\mathcal{P}^{\text{grd}}_{\text{grd}} \equiv \mathcal{P}^{x}_{\text{within}}$. We do not match ground and aerial images directly, i.e., $\mathcal{P}^{\text{grd}}_{\text{aerial}} = \emptyset$, as few accurate matches can be found and removing these pairs accelerate the feature matching process. Both $\mathcal{P}^{\text{trans}}_{\text{aerial}}$ and $\mathcal{P}^{\text{aerial}}_{\text{aerial}}$ are exhaustive. Finally, ground-transition and transition-transition pairings can be defined as:

$$\begin{aligned} \mathcal{P}^{\text{grd}}_{\text{trans}} &= \{(\mathcal{I}^{\text{grd}}_{i}, \mathcal{I}^{\text{trans}}_{j}) | i \in \mathcal{S}^{\text{grd}}_{\text{front}}, j \in \mathcal{S}^{\text{trans}}_{\text{front}}\} \cup \{(\mathcal{I}^{\text{grd}}_{i}, \mathcal{I}^{\text{trans}}_{j}) | i \in \mathcal{S}^{\text{grd}}_{\text{back}}, j \in \mathcal{S}^{\text{trans}}_{\text{back}}\}, \\ \mathcal{P}^{\text{trans}}_{\text{trans}} &= \{(\mathcal{I}^{\text{trans}}_{i}, \mathcal{I}^{\text{trans}}_{j}) | i \in \mathcal{S}^{\text{trans}}_{\text{front}}, j \in \mathcal{S}^{\text{trans}}_{\text{front}}\} \cup \{(\mathcal{I}^{\text{trans}}_{i}, \mathcal{I}^{\text{trans}}_{j}) | i \in \mathcal{S}^{\text{trans}}_{\text{back}}, j \in \mathcal{S}^{\text{trans}}_{\text{back}}\}. \end{aligned} \tag{3}$$

We use both SP+SG (DeTone et al., 2018; Sarlin et al., 2020) and RoMa (Edstedt et al., 2024b) to compute correspondences based on $\mathcal{P}^{\text{cross\_elev}}$, and COLMAP (Schönberger & Frahm, 2016) to perform SfM. Finally, we select the best results from different matchers.

3.3 Coordinate Alignment

Since SfM systems estimate camera up to an *arbitrary* scale and orientation, we need to align multiple coordinate systems together. Given the same 3D points in two coordinate systems, Procrustes Alignment (Gower, 1975) finds the scale, rotation, and translation $\{s, r, t\}$ transformations between them:

$$s^*, r^*, t^* = \arg\min_{s,r,t} \sum_i \|s(rp^i_{\mathcal{X}} + t) - p^i_{\mathcal{Y}}\|^2, \tag{4}$$

where $p^{i,\mathcal{X}}$ and $p^{i,\mathcal{Y}}$ are 3D points in coordinate system $\mathcal{X}$ and $\mathcal{Y}$. To better align the camera systems, we optimize based on both the camera center $P^{i,\mathcal{X}}_{\text{pos}} \in \mathbb{R}^{1\times 3}$ and rotation $R^{i,\mathcal{X}} \in \mathbb{R}^{3\times 3}$. Specifically, we represent camera rotation by backprojecting three points based on camera center and rotation:

$$P^{i,\mathcal{X}}_{\text{rot}} = P^{i,\mathcal{X}}_{\text{pos}} + s^{\mathcal{X}} R^{i,\mathcal{X}}, P^{i,\mathcal{Y}}_{\text{rot}} = P^{i,\mathcal{Y}}_{\text{pos}} + s^{\mathcal{Y}} R^{i,\mathcal{Y}} \tag{5}$$

where $s^{\mathcal{X}} = ||\sigma^{\mathcal{X}}||$, and $\sigma^{\mathcal{X}}$ is the standard deviations of $P^{i,\mathcal{X}}_{\text{pos}}$; $s^{\mathcal{X}}$ is similarly defined. Based on $P^{i,\mathcal{X}}_{\text{rot}}$ and $P^{i,\mathcal{Y}}_{\text{rot}}$, we update Eq. (4) as follows:

$$(s^*, r^*, t^*) = \arg\min_{s,r,t} \sum_i \|s(rP^{i,\mathcal{X}}_{\text{pos}} + t) - P^{i,\mathcal{Y}}_{\text{pos}}\|^2 + \|s(rP^{i,\mathcal{X}}_{\text{rot}} + t) - P^{i,\mathcal{Y}}_{\text{rot}}\|^2. \tag{6}$$

We show that this significantly improves the rotation alignment accuracy in our appendix.

Single Building Alignment. For every building, we obtain $(s^*_{\text{grd}}, r^*_{\text{grd}}, t^*_{\text{grd}})$ from the ground-only coordinate system to the cross-elevation coordinate system. This is done by applying Eq. (6) on the panorama sequence, which are calibrated in both systems. Similarly, we find $(s^*_{\text{aerial}}, r^*_{\text{aerial}}, t^*_{\text{aerial}})$ for aerial cameras based on the shared aerial sequence. All images of a single building can then be transformed into a unified coordinate frame.

Campus-wide Alignment To put cameras from all buildings into the same system, we perform a similar alignment process. To accomplish this, we first calibrate a subset of aerial images from every building, captured during summer from an altitude of 60m. Based on the shared aerial images, we use Eq. (6) to find the transformation of every building's individual coordinate system to the campus-wide aerial calibration.

4 Experiments

We examine SoTA camera calibration and dense reconstruction algorithms on ULTRA-360. Specifically, two challenges are explored in camera calibration: 1. finding true positive matches between far-apart images, e.g., across elevation; 2. avoiding false positive matches between images that are

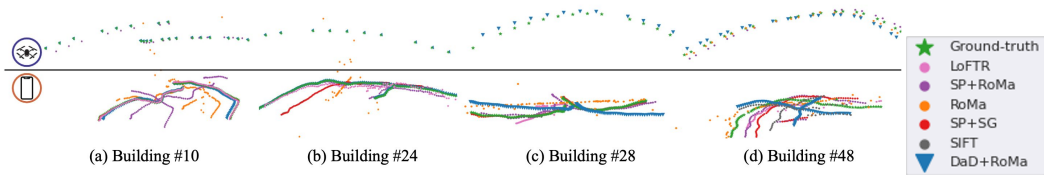
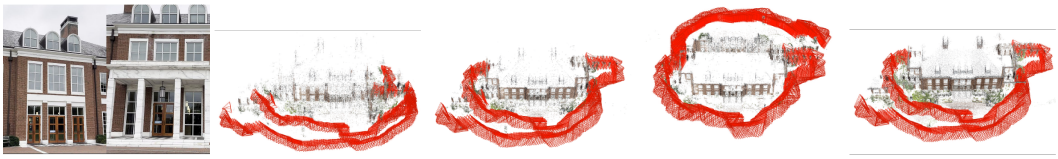


Figure 3: Visualization of multi-elevation camera poses obtained from different matching methods.

not visible to each other but have similar patterns. Two challenges are explored in dense reconstruction: 1. cross-elevation NVS and 2. multi-appearance NVS. Through experiments, we observe progress in these four challenges and many areas for future research to improve upon.

Cross-Elevation Feature Matching. For each building in ULTRA-360, we select a portion of the front side perspective ground images and aerial images acquired at 120m, *without* the transitional images that connect them. We test six popular feature matching algorithms: SIFT (Lowe, 2004), SP+SG (DeTone et al., 2018; Sarlin et al., 2020), SP+LG (DeTone et al., 2018; Lindenberger et al., 2023a), LoFTR (Sun et al., 2021), RoMa (Edstedt et al., 2024b), and RoMa filtered by two feature extractors, SuperPoint (DeTone et al., 2018) and DaD (Edstedt et al., 2025). Exhaustive matching is used for all scenarios mentioned above. In addition, we test four contemporary feed-forward matching methods: VGGSfM (Wang et al., 2024), VGGT (Wang et al., 2025), MASt3R (Leroy et al., 2024) and MASt3R-SfM (Duisterhof et al., 2025). We report AUC@10, computed from Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA). To isolate the cross-elevation challenge, AUC is computed only over ground–aerial pairs. For each ground–aerial pair, we measure the angular errors in rotation and translation and take the AUC of the minimum of RRA and RTA over 10-degrees threshold, a common metric for calibration.

As shown in Table. 3 and visualized in Fig. 3, calibrating cross-elevation images is challenging. In general, no algorithms can correctly calibrate all scenarios correctly. Interestingly, RoMa (Edstedt et al., 2024b)-based methods are the only ones with the ability to find cross-elevation correspondences. This can be attributed to its DINOv2 (Oquab et al., 2023) foundation model backbone. Despite the high sensitivity, RoMa (Edstedt et al., 2024b) is prone to false positives, as ground images are often falsely matched to each other due to similar patterns on the building. To this end, we find that SP (DeTone et al., 2018) or DaD (Edstedt et al., 2025) can help filter these false positives. However, they can still fail in Fig. 3(c) and (e).



(a) Front & Back Building #54 | (b) SP+SG Exhaustive | (c) SP+SG netVLAD | (d) SP+SG with DP++ | (e) DaD+RoMa with DP++

Figure 4: Visualization of calibration with various scene graph optimization methods given visual ambiguity in (a). All but (d) lead to suboptimal calibration solutions.

Automated Scene Graph Optimization. Various methods have been proposed to optimize the viewing scene graph to remove visually ambiguous pairs (Cai et al., 2023; Xiangli et al., 2024) and reduce excessive computation (Arandjelovic et al., 2018; Berton et al., 2023). These approaches are important for unconstrained calibration, where sensitive feature matchers are necessary, and false positive matches pose significant challenges. To this end, we evaluate several methods on ULTRA-360, particularly the ground panorama sequences.

As visualized in Fig. 4, exhaustive matching often leads to the worst results both in accuracy and computation due to visual ambiguities. NetVLAD (Arandjelovic et al., 2018) reduces computation by cutting down unnecessary pairs, but cannot resolve doppelgangers. Doppelganger++ (Xiangli et al., 2024) simplifies the scene graph and address doppelgangers to some degree; however, sensitive matchers like RoMa (Edstedt et al., 2024b) still finds enough false matches to lead to a deformed calibration, whereas SuperGlue (Sarlin et al., 2020) is less sensitive but more specific, achieving the correct solution. In summary, selecting appropriate scene graphs and feature matchers to obtain

Table 3: Cross-Elevation camera poses obtained from different matching methods. Measured in AUC@10 (higher is better).

Method	Building #10	#24	#28	#34	#48	#49	#54
LoFTR	0	0	0	0	0	0	0
SP+RoMa	0.3738	0	0	0	0.6986	0	0.5966
RoMa	0.0854	0.0023	0	0.0036	0.5030	0	0.1388
SP+SG	0	0	0	0	0	0	0
SP+LG	0	0	0	0	0	0	0
SIFT	0	0	0	0	0	0	0
DaD+RoMa	0.6941	0.8000	0	0.7915	0.5465	0.7440	0.6380
VGGT	0.1384	0	0	0	0.0003	0	0
VGGSfM	0	0	0	0	0	0	0
MASt3R	OOM	0	0	0	0	0	0
MASt3R-SfM	0	0	0	0	0	0	0

Table 4: Quantitative evaluation on multi-elevation reconstruction. We split the training set into either ground-only (G), aerial-only (A), or ground-aerial combined (GA) imagery. The test views are also separated into ground-only (G) and aerial-only (A) subsets. Due to different collection conditions, we only evaluate DSIM in *cross-elevation rendering*.

		Block-MERF			Splatfacto-W			CityGS V2			Scaffold-GS			Octree-GS			EVER		
Train	Test	PSNR	SSIM	DSIM	PSNR	SSIM	DSIM	PSNR	SSIM	DSIM	PSNR	SSIM	DSIM	PSNR	SSIM	DSIM	PSNR	SSIM	DSIM
G	G	21.020	0.609	0.118	21.925	0.657	0.166	20.702	0.655	0.168	21.551	0.658	0.122	21.360	**0.667**	**0.109**	**21.971**	0.641	0.146
A	G	-----	-----	0.588	-----	-----	0.639	-----	-----	**0.522**	-----	-----	0.595	-----	-----	0.608	-----	-----	0.619
GA	G	19.655	0.574	0.235	**21.569**	0.647	0.183	20.585	0.643	0.188	21.140	0.635	0.154	21.184	**0.653**	**0.116**	21.522	0.624	0.175
A	A	27.451	0.779	0.015	29.440	0.860	0.016	28.997	0.840	0.009	**30.286**	**0.878**	0.006	29.950	0.874	**0.005**	26.397	0.720	0.023
G	A	-----	-----	0.847	-----	-----	**0.714**	-----	-----	0.743	-----	-----	0.822	-----	-----	0.755	-----	-----	0.740
GA	A	13.453	0.106	0.407	23.206	0.669	0.042	20.129	0.598	0.173	26.135	0.748	**0.022**	**26.488**	**0.759**	0.024	23.433	0.644	0.039

good calibration still requires manual inspection and expertise. For more complete metrics and visualizations regarding image registration, please refer to our appendix.

Large-scale Dense Reconstruction and NVS. We select ten buildings from ULTRA-360 to evaluate current progress in robust, large-scale 3D reconstruction. For each building, we split training data into three configurations: 1. ground images only, 2. aerial images only, 3. both ground and aerial images. For each configuration, we evaluate from held-out ground and aerial cameras separately.

For baselines, we choose six SoTA methods for evaluation: Splatfacto-W (Xu et al., 2024a), Block-MERF (Song et al., 2024), CityGaussianV2 (Liu et al., 2024b), Scaffold-GS (Lu et al., 2024), Octree-GS (Ren et al., 2024) and EVER (Mai et al., 2024). Multiple metrics are used to evaluate NVS performance: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) are used for low-level quality evaluation. Perceptual similarity metrics DreamSim (Fu et al., 2023) (DSIM) are used to quantify semantic similarity, which helps in cases where pixel-wise groundtruth is not available due to e.g., changed lighting conditions.

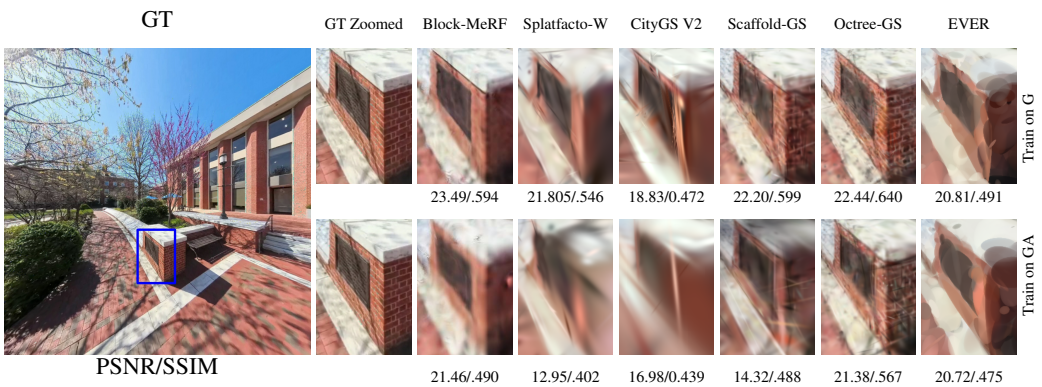


Figure 5: Visualization of ground image rendering from different reconstruction methods and two training configurations: ground-only images (G) and ground+aerial images (GA).

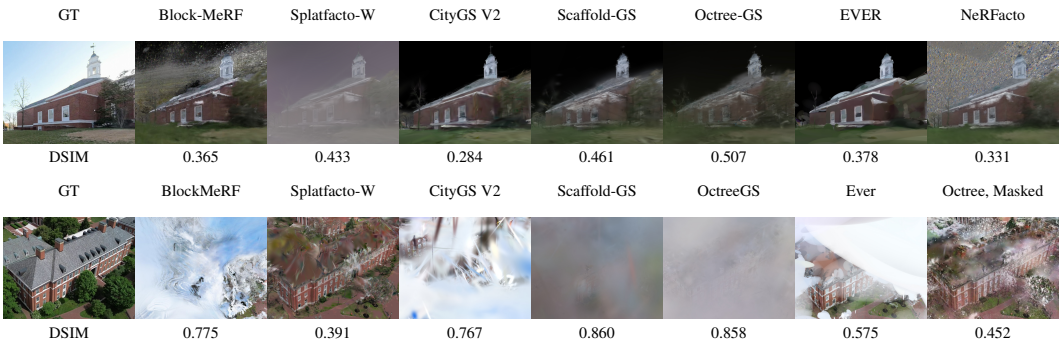


Figure 6: Top: ground view renderings from aerial-image-only reconstructions. Bottom: aerial view renderings from ground-image-only reconstructions.

As shown in Table 4, and visualized in Figures 5 and 6, we observe that Octree-GS (Ren et al., 2024) performs the best out of all methods, particularly when training data contain multi-elevation images. This can be attributed to its Level-of-Detail implementation. Scaffold-GS (Lu et al., 2024) achieves comparable fidelity through hierarchical Gaussian decomposition. All methods perform much worse given cross-elevation images for training compared to using single-elevation only. Interestingly, this may not be due to limited capacity. As shown in Table 5, cross-elevation reconstructions have significantly less Gaussians compared to single-elevation reconstructions, despite being strictly a superset in its training data. This likely indicates densification algorithms experience challenges when Gaussians' positional gradients are pulled from different directions.

By rendering cross-elevation cameras, we can observe various artifacts from current methods. Significant floaters exist when ground-only reconstructions are rendered from aerial perspectives. Notably, Splatfacto-W (Xu et al., 2024a) achieves superior aerial rendering through its background modeling. As shown in Figure 6, we also implement an implicit neural network to model sky in Octree-GS, which significantly reduce floaters. CityGS (Liu et al., 2024a;b) performs best on ground-view reconstruction from aerial data via geospatial-aware Gaussian priors optimized for large-scale aerial image.

Multi-Appearance Reconstruction and Zero-Shot NVS. ULTRA-360 contains multi-view sequences collected at different time. We use these sequences to evaluate multi-appearance reconstruction. Wild-GS (Xu et al., 2024b) and Gaussian-Wild (Zhang et al., 2024) are used as baselines, both of which require test image for evaluation. Unlike previous datasets (Snavely et al., 2006), ULTRA-360 has access to multi-view groundtruth at *every appearance*. This allows us to evaluate the effect of per-training-image embeddings on test images. As shown in in Table 6, we find that previous approaches lead to severe entanglement between view direction and the general appearances. Specifically, if we apply embedding from a training image that is the farthest away from the current test view, a significant drop in performance can be observed. The larger the performance drop suggests that the embeddings and networks are learned to overfit the input images, instead of the general 3D appearance. By modifying the per-image embedding to a time-based embedding, we can both remove the reliance on test-images at render time and achieve more 3D consistent appearance modeling. We provide more details and visualizations in our appendix.

Table 5: The average number of 3D Gaussians under different training configurations

Train	Splatfacto-W	CityGS V2	Octree-GS	EVER
G	340244	569325	3191058	535701
A	630093	287026	527991	70738
GA	309018	241688	2230053	262366

Table 6: Quantitative evaluation of multi-appearance reconstruction and rendering based on ULTRA-360.

	Wild-GS			Gaussian-Wild		
	PSNR	SSIM	DSIM	PSNR	SSIM	DSIM
Test Image Embedding	28.133	0.864	0.015	26.528	0.767	0.020
Nearest Train Image Embedding	28.003	0.863	0.014	26.567	0.757	0.020
Farthest Train Image Embedding	22.506	0.770	0.061	25.621	0.757	0.023
Time Embedding	27.973	0.860	0.014	26.277	0.762	0.021

5 Discussion and Conclusion

In this work, we propose a dataset called ULTRA-360 for Unconstrained Large-scale Temporal 3D Reconstruction across Altitudes. ULTRA-360 contains 37.7k frames collected from hundreds of

videos across the campus and includes academic buildings from multiple seasons, multiple elevations, and multiple camera types. To this end, we ensure cameras from different elevations can find correspondences based on ground-to-aerial transitional images. We also eliminate false matches through manually defined scene graphs.

Popular feature matching and scene graph optimization algorithms are evaluated to measure how imagery from ULTRA-360 can be calibrated without assistance. Some methods demonstrate significant improvement in finding difficult true positives, at the cost of more false positives. While proper filtering based on keypoint extraction can lead to less false positives, current camera calibration pipeline still fall into incorrect solutions due to visual ambiguities, even with scene graph optimization. This showcases the need for a potentially more global approach in addressing doppelgangers rather than relying pair-wise prediction.

We also evaluate various dense reconstruction methods on ULTRA-360. We find that current methods, even those designed for large scale reconstruction, perform much worse given cross-elevation images for training compared to using single-elevation only. This likely indicates limitations in densification algorithms at scale. Multi-appearance reconstruction is also benchmarked. Several methods require access to test-time images to model appearance. Based on ULTRA-360, we find that these methods tend to generate embeddings that are heavily over-fitted to specific viewpoint, leading to suboptimal results to other views of the same appearance.

ULTRA-360 provides many novel directions for research, including the study on out-of-distribution NVS, campus-scale immersive 4D reconstruction, and potentially serving as test-grounds for evaluating geometric plausibility for generative models. In the future, we will continue to expand on ULTRA-360 to include more buildings and temporal variations.

REFERENCES

Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1437–1451, 2018. doi: 10.1109/TPAMI.2017.2711011. URL https://doi.org/10.1109/TPAMI.2017.2711011.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5460–5469. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00539. URL https://doi.org/10.1109/CVPR52688.2022.00539.

Fabio Bellavia, Jiri Matas, Dmytro Mishkin, Luca Morelli, Fabio Remondino, Amy Tabb, Eduard Trulls, Kwang Moo Yi, Sohier Dane, Addison Howard, and María Cruz. Image matching challenge 2025. https://kaggle.com/competitions/image-matching-challenge-2025, 2025. Kaggle.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11080–11090, 2023.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 11618–11628. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01164. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html.

Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 34–44. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00010. URL https://doi.org/10.1109/ICCV51070.2023.00010.

Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 12933–12942. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01260. URL https://doi.org/10.1109/CVPR52688.2022.01260.

David J. Crandall, Andrew Owens, Noah Snavely, and Dan Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pp. 3001–3008. IEEE Computer Society, 2011. doi: 10.1109/CVPR.2011.5995626. URL https://doi.org/10.1109/CVPR.2011.5995626.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 224–236. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPRW.2018.00060. URL http://openaccess.thecvf.com/content_cvpr_2018_workshops/w9/html/DeTone_SuperPoint_Self-Supervised_Interest_CVPR_2018_paper.html.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *CoRR*, abs/2409.19152, 2024. doi: 10.48550/ARXIV.2409.19152. URL https://doi.org/10.48550/arXiv.2409.19152.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Mast3r-sfm: A fully-integrated solution for unconstrained structure-from-motion. In *International Conference on 3D Vision, 3DV 2025, Singapore, March 25-28, 2025*, pp. 1–10. IEEE, 2025. doi: 10.1109/3DV66043.2025.00008. URL https://doi.org/10.1109/3DV66043.2025.00008.

Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024a.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 19790–19800. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.01871. URL https://doi.org/10.1109/CVPR52733.2024.01871.

Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DaD: Distilled Reinforcement Learning for Diverse Keypoint Detection. *arXiv preprint arXiv:2503.07347*, 2025.

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. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9f09f316a3eaf59d9ced5ffaefe97e0f-Abstract-Conference.html.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: create anything in 3d with multi-view diffusion models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/89e4433fec4b99f1d859db57af1e0a0f-Abstract-Conference.html.

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *CoRR*, abs/2308.04079, 2023. doi: 10.48550/ARXIV.2308.04079. URL https://doi.org/10.48550/arXiv.2308.04079.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4):78:1–78:13, 2017. doi: 10.1145/3072959.3073599. URL https://doi.org/10.1145/3072959.3073599.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/25c0fe7b157821dd3140727dc07461da-Abstract-Conference.html.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *CoRR*, abs/2407.08447, 2024b. doi: 10.48550/ARXIV.2407.08447. URL https://doi.org/10.48550/arXiv.2407.08447.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXII*, volume 15130 of *Lecture Notes in Computer Science*, pp. 71–91. Springer, 2024. doi: 10.1007/978-3-031-73220-1_5. URL https://doi.org/10.1007/978-3-031-73220-1_5.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3182–3192. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00297. URL https://doi.org/10.1109/ICCV51070.2023.00297.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3292–3310, 2023. doi: 10.1109/TPAMI.2022.3179507. URL https://doi.org/10.1109/TPAMI.2022.3179507.

Huangjing Lin, Hao Chen, Qi Dou, Liansheng Wang, Jing Qin, and Pheng-Ann Heng. Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pp. 539–546. IEEE Computer Society, 2018. doi: 10.1109/WACV.2018.00065. URL https://doi.org/10.1109/WACV.2018.00065.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *CVPR*, 2024.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 17581–17592. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.01616. URL https://doi.org/10.1109/ICCV51070.2023.01616.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023b.

Yang Liu, He Guan, Chuanchen Luo, Lue Fan, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. *CoRR*, abs/2404.01133, 2024a. doi: 10.48550/ARXIV.2404.01133. URL https://doi.org/10.48550/arXiv.2404.01133.

Yang Liu, Chuanchen Luo, Zhongkai Mao, Junran Peng, and Zhaoxiang Zhang. Citygaussianv2: Efficient and geometrically accurate reconstruction for large-scale scenes. *CoRR*, abs/2411.00771, 2024b. doi: 10.48550/ARXIV.2411.00771. URL https://doi.org/10.48550/arXiv.2411.00771.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://doi.org/10.1023/B:VISI.0000029664.99615.94.

D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pp. 1150–1157 vol.2, 1999. doi: 10.1109/ICCV.1999.790410.

Chongshan Lu, Fukun Yin, Xin Chen, Wen Liu, Tao Chen, Gang Yu, and Jiayuan Fan. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 7523–7533. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00695. URL https://doi.org/10.1109/ICCV51070.2023.00695.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 20654–20664. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01952. URL https://doi.org/10.1109/CVPR52733.2024.01952.

Alexander Mai, Peter Hedman, George Kopanas, Dor Verbin, David Futschik, Qiangeng Xu, Falko Kuester, Jonathan T. Barron, and Yinda Zhang. EVER: exact volumetric ellipsoid rendering for real-time view synthesis. *CoRR*, abs/2410.01804, 2024. doi: 10.48550/ARXIV.2410.01804. URL https://doi.org/10.48550/arXiv.2410.01804.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 7210–7219. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00713. URL https://openaccess.thecvf.com/content/CVPR2021/html/Martin-Brualla_NeRF_in_the_Wild_Neural_Radiance_Fields_for_Unconstrained_Photo_CVPR_2021_paper.html.

Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023.

Zhenxing Mi and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PQ2zoIZqvm.

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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pp. 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_24. URL https://doi.org/10.1007/978-3-030-58452-8_24.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. doi: 10.48550/ARXIV.2304.07193. URL https://doi.org/10.48550/arXiv.2304.07193.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

Cheng Peng and Rama Chellappa. PDRF: progressively deblurring radiance field for fast scene reconstruction from blurry images. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 2029–2037. AAAI Press, 2023. doi: 10.1609/AAAI.V37I2.25295. URL https://doi.org/10.1609/aaai.v37i2.25295.

Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman. MERF: memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Trans. Graph.*, 42(4):89:1–89:12, 2023. doi: 10.1145/3592426. URL https://doi.org/10.1145/3592426.

Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *CoRR*, abs/2403.17898, 2024. doi: 10.48550/ARXIV.2403.17898. URL https://doi.org/10.48550/arXiv.2403.17898.

Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 20626–20636. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01976. URL https://doi.org/10.1109/CVPR52729.2023.01976.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 4937–4946. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00499. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Sarlin_SuperGlue_Learning_Feature_Matching_With_Graph_Neural_Networks_CVPR_2020_paper.html.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006. doi: 10.1145/1141911.1141964. URL https://doi.org/10.1145/1141911.1141964.

Kaiwen Song, Xiaoyi Zeng, Chenqu Ren, and Juyong Zhang. City-on-web: Real-time neural rendering of large-scale scenes on the web. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pp. 385–402. Springer, 2024. doi: 10.1007/978-3-031-72970-6_22. URL https://doi.org/10.1007/978-3-031-72970-6_22.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 8922–8931. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00881. URL https://openaccess.thecvf.com/content/CVPR2021/html/Sun_LoFTR_Detector-Free_Local_Feature_Matching_With_Transformers_CVPR_2021_paper.html.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8248–8258, 2022a.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8238–8248. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.00807. URL https://doi.org/10.1109/CVPR52688.2022.00807.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, Justin Kerr, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In Erik Brunvand, Alla Sheffer, and Michael Wimmer (eds.), *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pp. 72:1–72:12. ACM, 2023. doi: 10.1145/3588432.3591516. URL https://doi.org/10.1145/3588432.3591516.

Yutao Tang, Yuxiang Guo, Deming Li, and Cheng Peng. Spars3r: Semantic prior alignment and regularization for sparse 3d reconstruction. *CVPR*, 2025.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*, volume 15087 of *Lecture Notes in Computer Science*, pp. 197–214. Springer, 2024. doi: 10.1007/978-3-031-73397-0_12. URL https://doi.org/10.1007/978-3-031-73397-0_12.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly- throughs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 12912–12921. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01258. URL https://doi.org/10.1109/CVPR52688.2022.01258.

Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360° videos. *Advances in Neural Information Processing Systems*, 37:17743–17760, 2024.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotný. Vggsfm: Visual geometry grounded deep structure from motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21686–21697. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02049. URL https://doi.org/10.1109/CVPR52733.2024.02049.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00499. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 18074–18084. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01661. URL https://doi.org/10.1109/ICCV51070.2023.01661.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. *arXiv preprint arXiv: 2503.01774*, 2025.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, volume 13692 of *Lecture Notes in Computer Science*, pp. 106–122. Springer, 2022. doi: 10.1007/978-3-031-19824-3_7. URL https://doi.org/10.1007/978-3-031-19824-3_7.

Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features. *CoRR*, abs/2412.05826, 2024. doi: 10.48550/ARXIV.2412.05826. URL https://doi.org/10.48550/arXiv.2412.05826.

Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections. *CoRR*, abs/2407.12306, 2024a. doi: 10.48550/ARXIV.2407.12306. URL https://doi.org/10.48550/arXiv.2407.12306.

Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/bb11f79ad86f5e33e2a7c850cbdfed42-Abstract-Conference.html.

Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *CoRR*, abs/2406.10373, 2024c. doi: 10.48550/ARXIV.2406.10373. URL https://doi.org/10.48550/arXiv.2406.10373.

Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 8254–8263. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00798. URL https://doi.org/10.1109/CVPR52729.2023.00798.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 1787–1796. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00186. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yao_BlendedMVS_A_Large-Scale_Dataset_for_Generalized_Multi-View_Stereo_Networks_CVPR_2020_paper.html.

Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. URL https://github.com/autonomousvision/sdfstudio.

Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In Ales Leonardis,

Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXVI*, volume 15134 of *Lecture Notes in Computer Science*, pp. 341–359. Springer, 2024. doi: 10.1007/978-3-031-73116-7_20. URL https://doi.org/10.1007/978-3-031-73116-7_20.

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.

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting, 2023.

A Appendix / supplemental material

A.1 Reconstructed Campus Visualization

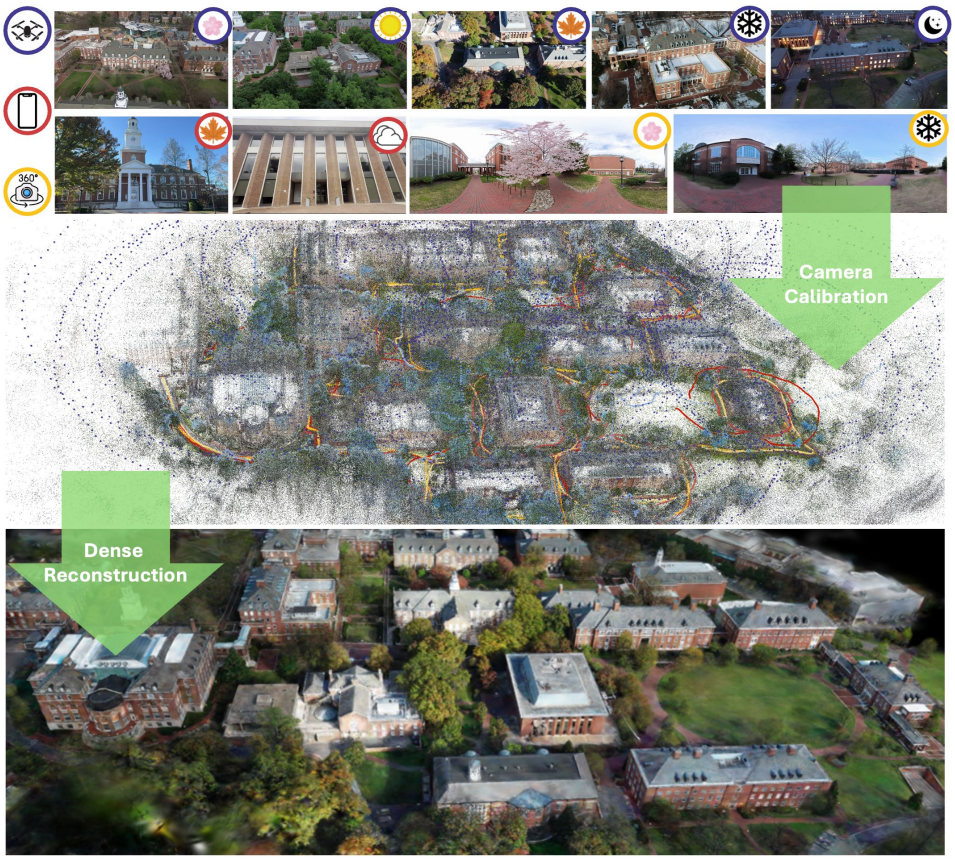


Figure A: A visualization of the reconstructed campus.

Figure A shows a visualization of the reconstructed campus based on our collected imagery over two years. The dataset is collected over multiple seasons, elevations, and multiple camera types to enable fully immersive 3D/4D reconstruction. All images have been calibrated into a unified coordinate system through a semi-automated process and manual verification.

A.2 Cross-Elevation Feature Matching

As shown in Figure B, we provide additional visualization of camera pose estimations for five buildings using six feature matching configurations, complementing the results shown in Figure 3. Overall, DaD+RoMa achieves higher accuracy, successfully estimating more camera poses with lower error. However, it fails to register the ground-level images in Figure B(a) and encounters false positive matches in Figure B(c), demonstrating the challenge of cross-elevation feature matching and underscoring the necessity of adopting the proposed single elevation calibration strategy.

We also visualize the absolute error of each estimated camera pose with respect to the ground truth after alignment in Figure C. Specifically, we sort these errors in ascending order; for images that fail to be calibrated, we assign a large error. DaD+RoMa is generally capable of estimating most camera poses except Figure C(d). Although SIFT struggles to register the multi-elevation images simultaneously, the successfully estimated poses tend to exhibit lower error, indicating higher confidence. We also observe that RoMa without any feature extractor leads to unstable results, which is reflected in the gradual increase in error across its estimated poses. This correlates with the observation that RoMa's raw correspondences contain both many true positives and false positives. In general, LoFTR and SP+SG performs similarly, compromising between sensitivity and specificity.

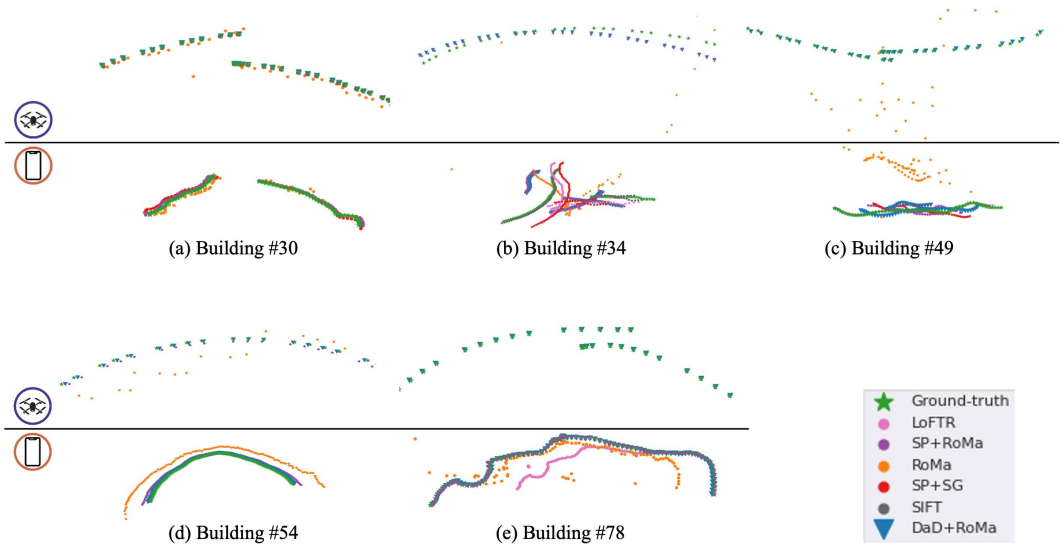


Figure B: Additional visualization of multi-elevation camera positions obtained from different matching methods.

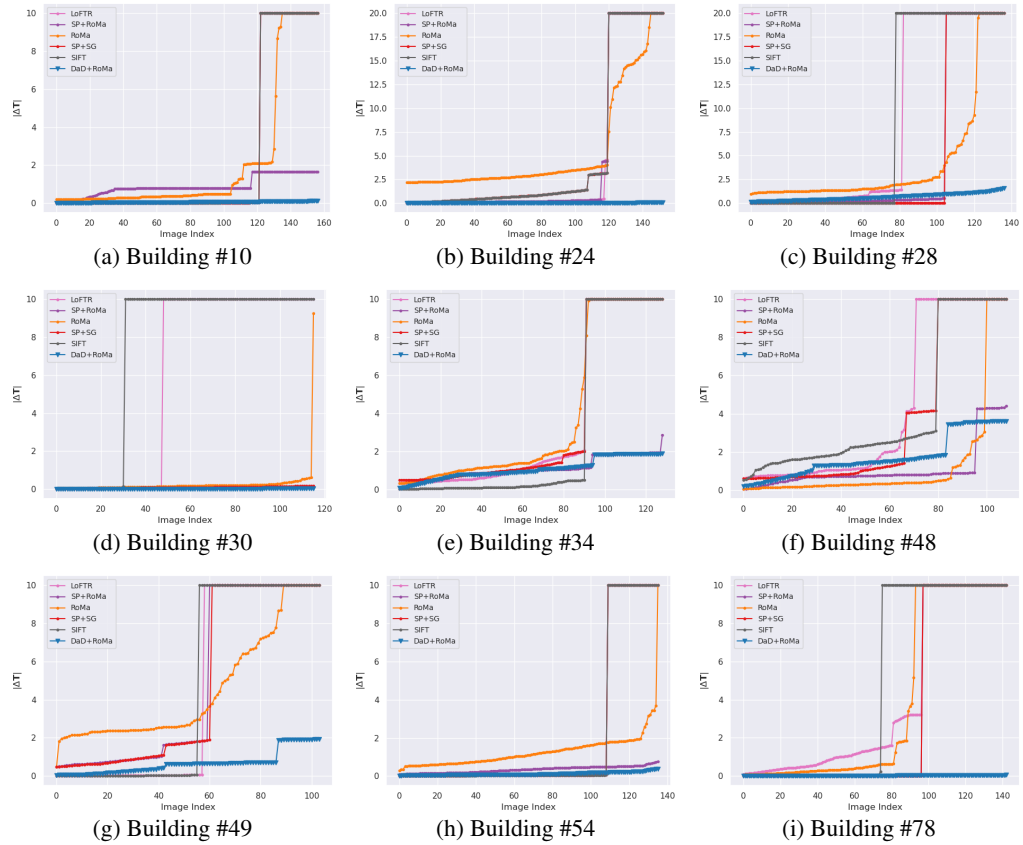


Figure C: Visualization of multi-elevation camera position error across nine buildings

A.3 Automated Scene Graph Optimization

We provide more examples in Figure D to demonstrate the challenge in visual ambiguities. Many buildings look similar from different angles. Exhaustive matching, e.g., with SP+SG, often fails. Without any knowledge of acquisition time, netVLAD (Arandjelovic et al., 2018) sometimes can

help prune away unnecessarily pairs to achieve better reconstruction; however, it's also very unreliable. Doppelganger++ (Xiangli et al., 2024) does better at eliminating confusing pairs, but different feature matchers can still be prone to errors in different scenarios.

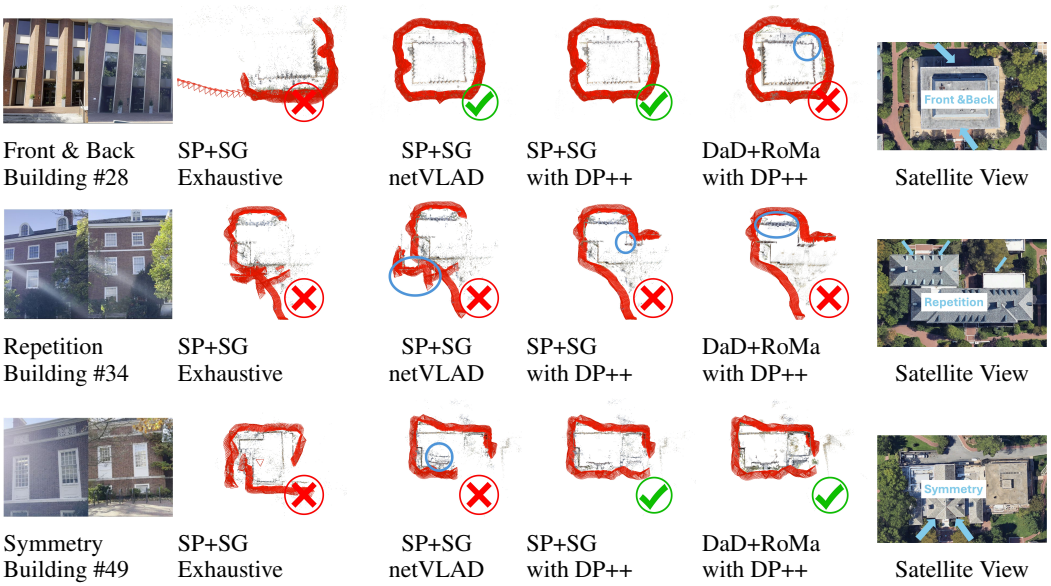


Figure D: Visualization of calibration with various scene graph optimization methods given visual ambiguity. A satellite view is provided to demonstrate the true locations of the images.

A.4 Large-scale Dense Reconstruction and NVS

Our dense reconstruction evaluation uses PSNR, SSIM, and DSIM as primary metrics. As a perceptual similarity metric LPIPS (Zhang et al., 2018) is also included in Appendix Tables A and B for completeness.

Table A: LPIPS on multi-elevation reconstruction.

Train	Test	Block-MERF	Splatfacto-W	CityGS V2	Scaffold-GS	Octree-GS	EVER
G	G	0.513	0.522	0.512	0.483	**0.443**	0.467
A	G	0.899	0.846	0.861	0.871	0.881	**0.829**
GA	G	0.602	0.539	0.553	0.541	**0.487**	0.503
A	A	0.175	0.188	0.173	**0.102**	0.123	0.299
G	A	0.920	0.881	0.912	0.911	**0.846**	0.869
GA	A	0.708	0.394	0.532	0.277	**0.266**	0.355

Table B: LPIPS of multi-appearance reconstruction.

	Wild-GS	Gaussian-Wild
Test Image Embedding	0.114	0.289
Nearest Train Image Embedding	0.115	0.288
Farthest Train Image Embedding	0.195	0.299
Time Embedding	0.118	0.298

A.5 Multi-Appearance Reconstruction and Zero-Shot NVS

Figure E shows the rendering results using different embeddings in the multi-appearance experiment. The difference maps between the rendered and ground truth images are also shown. It can be seen that the image rendered with the embedding farthest from the training view exhibits a significant overall appearance difference. These visual comparisons highlight a key drawback of per-image embeddings that they are view-dependent and lack consistency across different views.

A.6 Coordinate Alignment

Mip-NeRF 360	$E_R(\mu)\downarrow$	$E_T(\mu)\downarrow$
Procrustes Alignment	0.196	0.0144
+ RANSAC	0.179	**0.0114**
+ *Rotation Points*	**0.156**	0.0117

Table C: Improvements over Procrustes Alignment baseline in average rotation error E_R and translation error E_T. Incorporating rotation points further minimizes the overall error.

We test the alignment algorithm on the Mip-NeRF 360 (Barron et al., 2022) dataset. Specifically, we calibrate a sparse subset of the images, then attempt to align it to the groundtruth coordinate

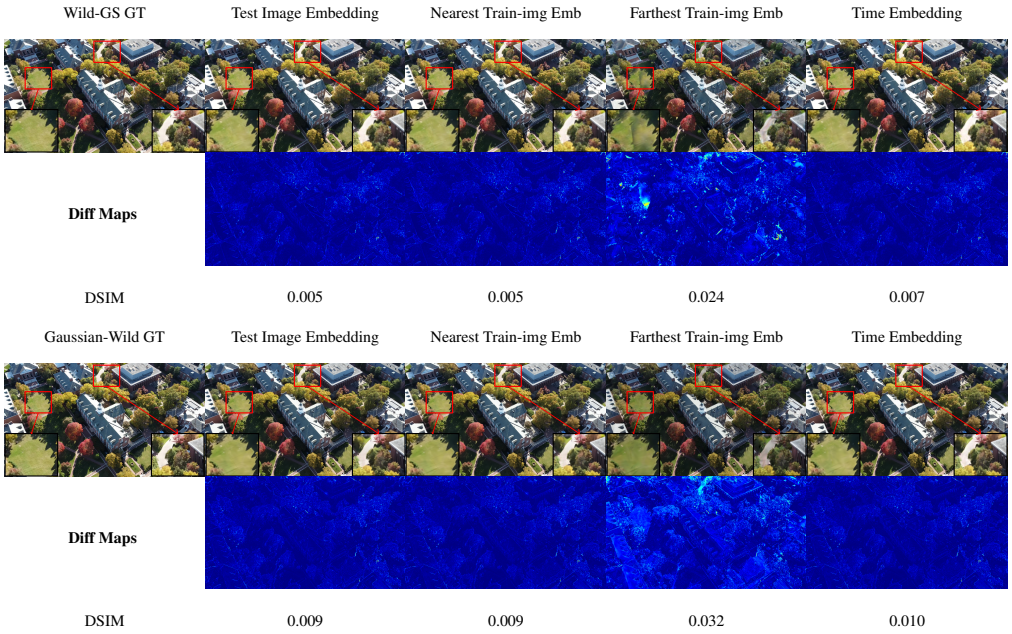


Figure E: The rendering result of Wild-GS and Gaussian-Wild on different appearance embeddings. Zoom-in images are shown in the bottom left and right; better viewed when magnified.

system. Sparse calibration leads to inaccuracy, and makes the alignment process more noisy. As shown in Tab. Table C, applying constraint on both the translation and rotation points indeed reduce the rotation error significantly.

A.7 Additional Scene Reconstruction Visualization

Please refer to the videos for additional rendering of 3D structures of the campus buildings.

A.8 The Use of Large Language Models (LLMs)

We do not use any large language models in this work when constructing our dataset nor when drafting the paper.