Scene Grounding In the Wild

Tamir Cohen[1] Leo Segre[1] Shay Shomer-Chai[1] Shai Avidan[1] Hadar Averbuch-Elor[2]
[1]Tel Aviv University [2]Cornell University

https://tau-vailab.github.io/SceneGround/

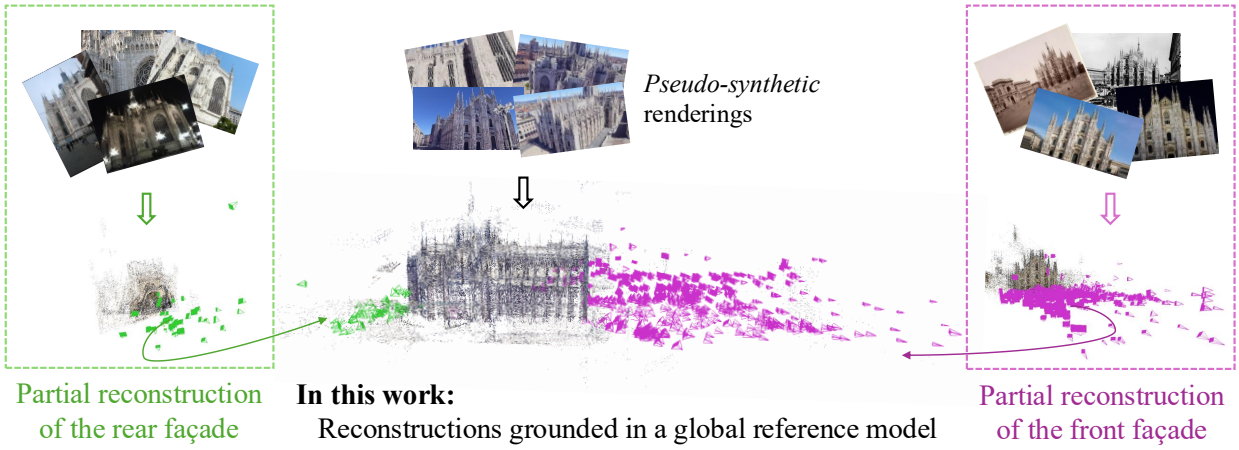

Figure 1. Given a partial 3D reconstruction produced by running structure from motion on Internet images capturing large-scale landmarks, such as the **front** or the **rear** façade of the Milan Cathedral depicted above, we present a technique for grounding this reconstruction in a complete 3D reference model of the scene. Reference models are constructed from *pseudo-synthetic* renderings extracted from Google Earth Studio. As illustrated above, our approach allows for merging partial, disjoint 3D reconstructions into a unified model.

Abstract

Reconstructing accurate 3D models of large-scale real-world scenes from unstructured, in-the-wild imagery remains a core challenge in computer vision, especially when the input views have little or no overlap. In such cases, existing reconstruction pipelines often produce multiple disconnected partial reconstructions or erroneously merge non-overlapping regions into overlapping geometry. In this work, we propose a framework that grounds each partial reconstruction to a complete reference model of the scene, enabling globally consistent alignment even in the absence of visual overlap. We obtain reference models from dense, geospatially accurate pseudo-synthetic renderings derived from Google Earth Studio. These renderings provide full scene coverage but differ substantially in appearance from real-world photographs. Our key insight is that, despite this significant domain gap, both domains share the same underlying scene semantics. We represent the reference model using 3D Gaussian Splatting, augmenting each Gaussian with semantic features, and formulate alignment as an inverse feature-based optimization scheme that estimates a global 6DoF pose and scale while keeping the reference model fixed. Furthermore, we introduce the WikiEarth *dataset, which registers existing partial 3D reconstructions with pseudo-synthetic reference models. We demonstrate that our approach consistently improves global alignment when initialized with various classical and learning-based pipelines, while mitigating failure modes of state-of-the-art end-to-end models. All code and data will be released.*

1. Introduction

One of the grand challenges in computer vision is to reconstruct the geometry of a 3D scene from an unstructured set of photographs. Over the past several decades, remarkable progress - ranging from classical structure-from-motion (SfM) pipelines [44, 48] to modern learning-based methods [27, 49, 56, 58] - has enabled increasingly accurate and detailed reconstructions, even from crowd-sourced

image collections containing transient objects, varying illumination, and significant appearance changes. However, despite these advances, 3D reconstruction frameworks fundamentally rely on sufficient visual overlap between input views to establish reliable geometric correspondences. This requirement is often violated in large-scale real-world image collections, where images are heavily biased toward a sparse set of iconic viewpoints. For example, as illustrated in Figure 1, tourists photographing the Milan Cathedral overwhelmingly capture its main entrance, with a smaller set of images depicting the rear façade. Such viewpoint bias yields multiple disconnected partial reconstructions - or worse, introduces erroneous geometry by collapsing non-overlapping observations into overlapping regions.

But what if we had access to a dense "oracle" reference model of the entire scene which could serve as a common anchor for all partial reconstructions? While crowd-sourced imagery rarely covers every region of interest, such a model can, in fact, be readily constructed from complementary sources of visual data. For example, tools like Google Earth Studio[1] can render dense, geospatially accurate views of real-world landmarks from arbitrary camera poses, yielding complete scene coverage. However, these renderings, previously referred to as pseudo-synthetic images [53], are generated from textured 3D meshes and therefore differ substantially in appearance from real-world crowd-sourced images. This pronounced domain gap makes it unclear how such reference models can be leveraged to align and unify partial reconstructions into a global coordinate system.

In this work, we introduce a technique that grounds partial reconstructions captured in the wild to a complete pseudo-synthetic–based reference model of the scene, effectively bridging the large appearance gap between the two domains. Our approach is motivated by the key observation that, despite substantial visual variation - both between crowd-sourced images and pseudo-synthetic renderings, and among the images themselves - all observations capture the same underlying scene *semantics*. We represent the reference model using 3D Gaussian Splatting (3DGS) [22] and cast the alignment as an inverse optimization problem [46, 64] that estimates a global transformation for each partial reconstruction while keeping the reference model fixed. In contrast to prior work that utilize a standard photometric loss for aligning input and rendered views, we propose a semantic feature-based robust optimization scheme that operates reliably on *real-world* image collections, even in the presence of outlier images that contain significant occlusions.

To evaluate our approach, we introduce the *WikiEarth* benchmark, pairing pseudo-synthetic reference models with thirty existing 3D reconstructions from WikiScenes [61]. We apply our inverse optimization scheme on top of various initializations spanning classical and learning-based pipelines, and observe consistent improvements in global alignment across multiple metrics. Additionally, we assess state-of-the-art feed-forward 3D models including DUSt3R [58], MASt3R [27], π^3 [60] and VGGT [56]. Our evaluation shows that despite strong performance in other settings, on our benchmark they frequently collapse non-overlapping partial reconstructions into incorrect geometries, highlighting the need for an external reference model and our semantic-based alignment approach. Finally, we demonstrate that our approach generalizes to reference models built from additional data sources, such as unstructured frames from drone videos. Our contributions include:

- A semantic alignment framework that grounds fragmented, non-overlapping partial reconstructions to a complete pseudo-synthetic–based reference model, effectively bridging large domain gaps between real-world imagery and rendered views.
- A robust feature–based optimization scheme, formulated as an inverse optimization problem over a 3D Gaussian Splatting representation, that replaces photometric cues with semantic features to achieve reliable alignment in challenging, in-the-wild conditions.
- The *WikiEarth* benchmark, which pairs pseudo-synthetic reference models with real-world 3D reconstructions, enabling systematic evaluation of cross-domain 3D alignment techniques.

2. Related Work

2.1. Sparse 3D Reconstruction

The goal of the sparse 3D reconstruction task is to reconstruct a scene given a sparse set of views, with little to no overlap between the images. Numerous works have been proposed for performing object-level sparse reconstruction [15, 30, 55, 65]. Much fewer works address this problem at scene-scale. In particular, Chen et al. [6] propose to learn discrete distributions over the 5D pose space.

Closely related to our problem setting, several prior work aim to reconstruct large-scale scenes that contain non-overlapping regions. Martin et al. [34] reconstruct indoor scenes given annotated floorplan maps. Cohen et al. [9] propose a technique for aligning indoor and outdoor reconstruction using windows detection. By contrast to these, in our work we are interested in leveraging accessible reference model for connecting partial reconstructions obtained from large photo collections. Another line of work has focused on the task of predicting 3D rotations (without estimating the relative translations) between non-overlapping images [2, 3, 11]. These methods utilize only a pair of images, rather than a full collection, and haven't been explored in the context of Internet imagery.

Recently, transformer-based 3D reconstruction meth-

[1] https://www.google.com/earth/studio/

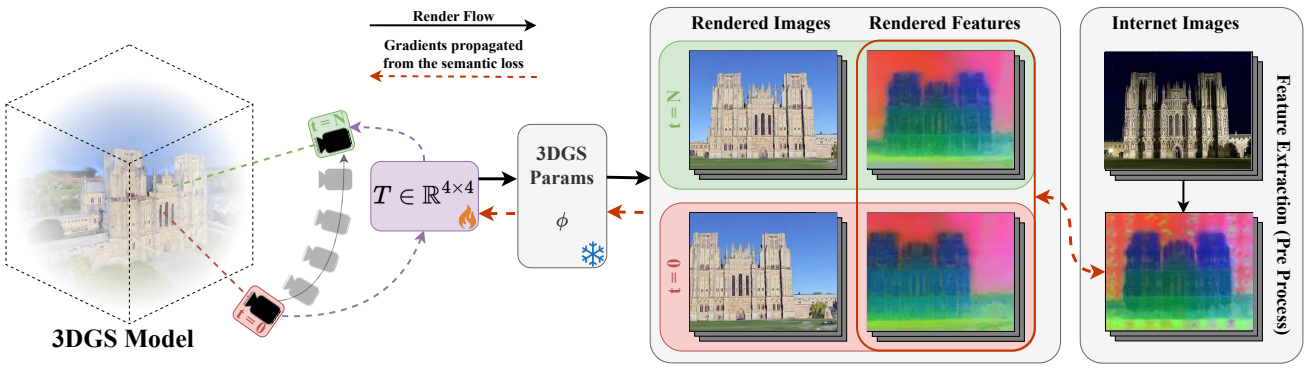

Figure 2. **Scene Grounding via Semantic Feature-based Robust Optimization**. Given a 3DGS reference model $\mathcal{M}$ (left) and a set of Internet images $\mathcal{I}$ (right), we propose an inverse optimization scheme that predicts a global 6DoF+scale alignment T while keeping the parameters of $\mathcal{M}$ fixed. We obtain an initial transformation T (in red) using a traditional SfM technique. During optimization, we calculate a semantic feature loss L_{sem} and backpropagate it to update T (converging to the rendered view in green after N steps).

ods have gained popularity. A series of works including DUSt3R [58], MASt3R [27], FASt3R [63], and Spann3r [54] have proposed using transformers to directly reconstruct sparse internet image collections. More recently, VGGT [56] introduced a large feed-forward transformer that can predict all key 3D attributes of a scene given input images. While these methods have achieved significant improvements in 3D reconstruction, these models are mostly trained on image sets with significant overlap, which limits their ability to handle sparse, non-overlapping collections. Additionally, memory constraints limit the number of input images that can be simultaneously processed. As we demonstrate in our paper, these methods cannot yet handle our challenging non-overlapping problem setting.

2.2. 3D Scene Registration

Registration has been studied extensively in the field of Computer Vision. Here we cover just the work most closely related to our work. In case the two 3D scenes are represented as point clouds, then a classical algorithm such as iterative closest poin t [1] and its many derivatives can be used. There are classic global methods that use 3D descriptors to aid matching [18] and then use a sparse subset of them for global alignment [68]. When the scene is represented as a mesh, mesh-based localization methods [38, 39] perform alignment by matching sparse features. Recent works learn the alignment features utilizing deep neural networks [8, 20, 43, 59, 66].

One method that utilizes it for NeRF registration is DReg-NeRF [7], which converts the NeRF to a voxel grid representation and trains a deep neural network for registration task. It achieves improved results over point cloud registration methods but requires a large training set. NeRF2NeRF [16] registers two 3D scenes, where both scenes are represented as a NeRF. They show an improvement over point cloud registration methods, but require user input at the initialization, which is not needed in our approach. GaussReg [4] also addresses the task of registering two 3D scene, however it operates on scenes represented as 3DGS. Later, VF-NeRF [46] extended NeRF to include Viewshed Fields, that capture visibility constraints with NeRF, and used it to register two NeRFs.

While point cloud, mesh-based and NeRF localization methods are related to our work, we address a different setting. Specifically, our approach operates within a framework that reconstructs a scene using a Gaussian Splatting base model from low-quality images. A more closely related work is iNeRF [64], which registers images to a NeRF model. Their method is based on back-propagating the photometric loss through the NeRF weights to optimize the six parameters of each camera pose, defining an SE3 transformation with an exponential parameterization. By contrast, our approach optimizes a global transformation aligning a partial reconstruction obtained by a SfM technique to a reference model, represented using 3DGS.

2.3. Semantic 3D Neural Representations

With the recent rise of 3D neural representations, various work has explored the problem of embedding semantic features over these 3D representation. These embedding are primarily used for tasks like semantic segmentation [52, 57, 67], object localization [5, 13, 19, 33, 36] and object recognition [17, 21]. For example LERF [24] augment NeRFs [35] with CLIP [41] embeddings, predicting a semantic feature field alongside the scene's geometry. Several recent works embed semantic features on a 3D Gaussian Splatting representation [40, 47, 69]. In particular, Feature 3DGS [69] proposed a method for distilling 3D feature fields from any 2D foundation models. They demonstrated distillation of LSeg [28] and SAM [25]. Several methods

specifically focus on learning semantic features for large-scale scenes [14, 26], such as the ones explored in our work. More closely related to our setting, Pixel-Perfect SfM [31] also minimizes a feature-metric loss. They optimize it as a part of standard SfM pipeline during bundle adjustment. By contrast, our approach uses semantic feature embeddings for aligning a 3DGS representation with a set of images.

3. Method

Our goal is to globally align a set of real-world images to a reference model. We assume the images were previously bundled together using a structure-from-motion technique (*e.g.*, COLMAP [45]), and treat them as a single meta-image $\mathcal{I}$. Our reference model is built from pseudo-synthetic rendered images, rendered from a mesh model such as the freely accessible[2] Google Earth Studio models, which, despite their low quality, provide extensive scene coverage. Specifically, we seek the 6DoF+scale transformation T to align the meta-image $\mathcal{I}$ with the reference model. The challenge lies in aligning these images, that are captured in uncontrolled environments with variable lighting, viewpoints, and occlusions, to a low-quality yet globally consistent reference model.

We frame the alignment problem as an inverse optimization problem as proposed in iNeRF [64]. However, unlike prior work, we iteratively refine a *global* 6DoF+scale transformation, leveraging information from multiple views to ground the meta-image $\mathcal{I}$ in the reference model $\mathcal{M}$. Furthermore, our reference model is a 3D Gaussian Splatting (3DGS) [23] model. 3DGS representations achieve real-time rendering speed, significantly outperforming NeRF-based methods, and hence are much better suited for inverse optimization-based solutions. Finally, to align meta-image $\mathcal{I}$, composed of images captured *in the wild*, with a reference model $\mathcal{M}$, we propose a semantic feature-based robust optimization scheme, as further detailed below.

Registration: We use an inverse-based approach for registration, optimizing the camera location based on rendering results (see Fig. 2). We freeze the 3DGS model parameters and introduce a 7-parameter vector: the first 6 represent rigid transformation in $SE3$, and the last parameter represents scale. Our method optimizes using semantic features as the objective function and addresses outliers with robust optimization, as detailed in the following paragraphs

Semantic Features: Prior inverse optimization-based techniques [46, 64] utilize a standard photometric loss for aligning the input images with views rendered from the neural 3D representation. In our problem setting, not only do the input images vary in appearance - for instance, due to illumination conditions and transient occlusions - but they also significantly differ from the views rendered from a (possibly) low-quality reference model. Consider the image pair in Figure 2 (top center). A standard color loss would not provide meaningful supervision for correctly aligning the image set to the reference model. Our key insight is that the underlying *semantics* is shared across the different scene observations, and thus semantic features can effectively guide the optimization procedure (Figure 2).

Specifically, we distill DINOv2 [37] features on the 3DGS model, where each Gaussian has both color and feature vectors, similar to the distillation approach used in Feature 3DGS [69]. During optimization, we use a L1 loss on these high-dimensional features, denoted as L_{sem}. DINO features have been previously utilized for various tasks, such as semantic image matching and semantic scene segmentation [37]. In our experiments, we show that they outperform other representations for our scene alignment task.

Robust Optimization: Even with our semantic loss, naively optimizing the transformation T between the meta-image $\mathcal{I}$ and the 3DGS model $\mathcal{M}$ fails because of outliers. See, for example, the image pairs illustrated in the bottom row of Figure 3. As shown in the figure, rendered views may appear behind floaters in the reference model (left example), and real-world images may contain occlusions (right example). These outliers increase the loss and impede the optimization's convergence.

To handle these outliers, we use robust optimization:

$$\hat{T} = \arg\min_{T} \varphi(\mathcal{L}(T|\mathcal{I}, \mathcal{M})), \tag{1}$$

where $\mathcal{L}$ is a summation of L_{sem} values, considering all the images in $\mathcal{I}$, and φ represents a robust loss function. In particular, we use the method of Least Trimmed Squares (LTS) [42]. In each optimization iteration we ignore the images with L_{sem} values larger than the median loss of the previous iteration.

Initialization: We evaluate various initialization methods for global alignment between the meta-image $\mathcal{I}$ and the reference model $\mathcal{M}$. Specifically, we initialize our method with COLMAP [45], gDLS+++ which finds the global 6Dof+scale meta-image transform by treating it as a single distributed camera [50], and the combination of SuperPoint [12] as the feature extractor and LightGlue [32] as the feature matcher (henceforth denoted as SP+LG). Additional details for these initialization methods are provided in the supplementary material.

Global Alignment and Model Assembly: Our proposed semantic feature-based optimization scheme is applied iteratively across all meta-images, gradually aligning each to the reference model. Each time, we independently align one meta-image to the model. This "puzzle-solving" process results in a cohesive, large-scale scene model that combines

[2] For non profit research purposes, as further detailed on their website.

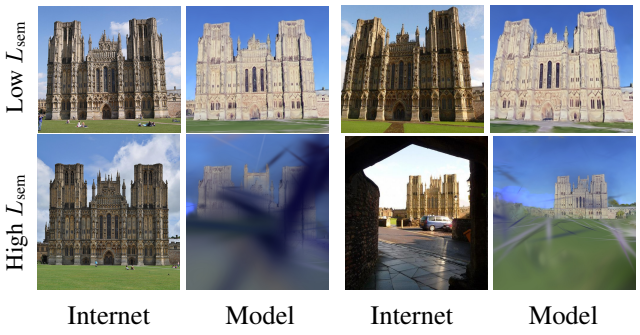

Figure 3. **Challenges of aligning internet photos to the reference model.** Visualization of input Internet images (first and third columns) and views rendered from the reference model at the ground-truth locations (second and fourth columns). As illustrated above, high L_{sem} values (bottom row) often indicate outlier images, which our approach overcomes via a robust optimization scheme, as further detailed in Section 3.

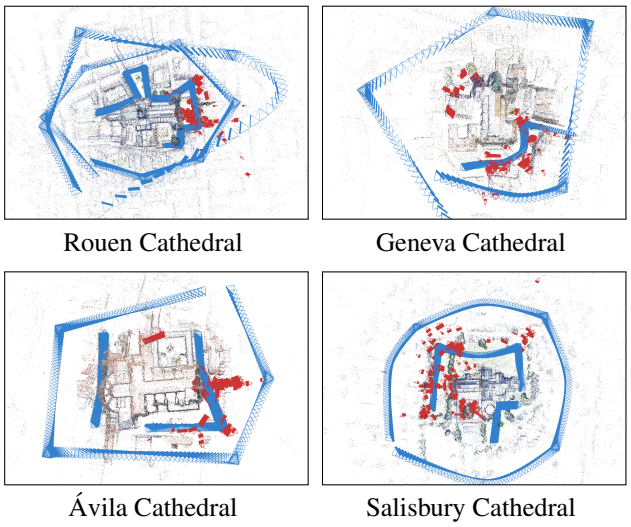

Figure 4. **The *WikiEarth* Benchmark**. Reconstruction of four landmarks from *WikiEarth*. The blue frustums depicts the rendered images from Google Earth Studio, and the red frustums the images from WikiScenes.

the partial reconstructions from each meta-image into a unified whole, overcoming limitations seen in traditional SfM methods that produce disjoint or incomplete reconstructions (as illustrated in Figure 1).

4. The *WikiEarth* Benchmark

To evaluate methods for grounding partial 3D reconstructions to complete scene models, we introduce the *WikiEarth* benchmark. While large collections of partial in-the-wild reconstructions exist (e.g., WikiScenes [61], MegaDepth [29], MegaScenes [51]), and many landmarks also have 3D models, there is no standardized dataset that provides ground-truth correspondences or alignments between the two. Moreover, existing in-the-wild reconstructions typically cover only limited portions of a scene, making accurate evaluation particularly challenging. Our benchmark fills this gap by supplying precise alignments between partial reconstructions and full scene geometry, enabling fair and reproducible quantitative comparison across alignment methods.

We augment 3D reconstructions from the WikiScenes dataset (meta-images) with reference models derived from Google Earth Studio. Google Earth studio is an animation tool for Google Earth's satellite and 3D imagery, which can be used for research purposes.

To create the reference model we first render camera trajectories from Google Earth Studio, while ensuring sufficient scene coverage. To achieve this, we render both aerial and ground level camera trajectories. An example of these camera trajectories is shown in Figure 4. The reference model is constructed by applying COLMAP to the rendered images of the landmark from Google Earth Studio, while utilizing the GPS coordinates of the rendered images.

We create ground-truth alignments between the WikiScenes meta-images and the Google Earth reference models through a fully supervised process. Specifically, we apply COLMAP with manually selected parameters to obtain an initial registration (exact parameters per scene are listed in the supplementary). We then visually inspect and filter out any misaligned images, retaining only those with high-quality alignment. A meta-image is included in our benchmark only if at least four of its images are successfully aligned to the reference model. Sample alignments are shown in the supplementary material.

The resulting *WikiEarth* benchmark consists of 32 different meta-images across 23 scenes from the WikiScenes dataset. On average, each meta-image contains 97 images, with a maximum of 713 and a minimum of 8. Additional statistics are available in the supplementary.

Figure 4 illustrates the reconstructions within the *WikiEarth* benchmark, where the blue frustums depict images rendered from Google Earth Studio and the red frustums Internet images from WikiScenes.

5. Experiments

In this section, we present our main results and comparisons. We compare performance to the COLMAP [45], GDLS+++ [50] and SP+LG [12, 32] baselines in Section 5.2. These baselines also serve as our initializations, as further detailed in Section 3. Additionally, we compare against recent feed-forward 3D models (specifically DUSt3R [58], MASt3R [27], π^3 [60], and VGGT [56]) in Section 5.3. Implementation details, additional experiments and interactive visualizations are provided in the supplementary material.

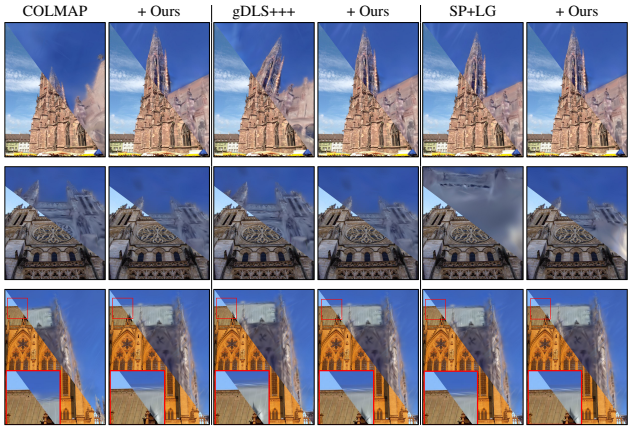

Figure 5. **Qualitative Comparison.** A visualization of the alignment results for our method compared to the three baselines. Each image shows the ground truth in the lower half and the rendered image from the reference model $\mathcal{M}$ after alignment in the top half. As demonstrated, our inverse optimization-based approach predicts precise transformations, even in the presence of challenging, inaccurate initializations.

5.1. Evaluation Metrics

We evaluate the predicted meta-image transformations on the *WikiEarth* benchmark using $\Delta T_{\mathcal{I}}$ and $\Delta R_{\mathcal{I}}$, which denote the average RMS translation and rotation errors of the registered images belonging to each meta-image $\mathcal{I}$. $\Delta R_{\mathcal{I}}$ is measured in degrees and $\Delta T_{\mathcal{I}}$ is expressed in the scale of the scene. All scenes are similarly scaled by COLMAP, using the rendered cameras as the reference scale.

We report ΔT and ΔR, which denote the average errors over all the meta-images (performance breakdown per meta-image is reported in the supplementary). Averages are computed only over successful meta-images, for which the method outputs a transformation.

In addition to the average errors, we report the meta-image transformation accuracy (MTA) and the percentage of outliers $O\%$, similarly to iNeRF [64]. MTA and $O\%$ are reported over predefined ratios. That is, a meta-image transformation is considered accurate if $\Delta R < 5°$ and $\Delta T < 0.2$, and categorized as an outlier if $\Delta R > 10°$ or $\Delta T > 0.5$. In the supplementary, we report MTA and $O\%$ over additional threshold values.

5.2. Comparison to Baselines

Quantitative results are reported in Table 1. As illustrated, our approach outperforms all baselines in all metrics and consistently improves the initialization performance. Performance breakdown over all scenes and configurations are provided in the supplementary material.

We also provide a qualitative comparison over different initializations in Figure 5. In Figure 6, we present alignment results of multiple meta-images across several landmarks, comparing our method and the COLMAP baseline to the ground truth. As can be observed from these visualizations, our method can successfully align meta-images with significantly erroneous initializations.

Table 1. **Comparison with Baselines**. We compare performance over the *WikiEarth* benchmark against multiple baselines. We report performance of each baseline (top rows), and our method initialized with the baselines (bottom rows). #Failures denotes the number of meta-images for which the initialization method failed to produce an alignment; errors are computed only over successful instances. As shown, our approach can be paired with a range of initializations, yielding significant improvements in most cases.

Methods	$\Delta R \downarrow$	$\Delta T \downarrow$	MTA$\uparrow$	$O\% \downarrow$	#Failures
gDLS+++	2.86	**0.12**	78	6	1/32
Ours (gDLS+++ init)	**2.69**	0.13	**84**	**3**	-
SP+LG	3.74	0.25	74	15	5/32
Ours (SP+LG init)	**3.13**	**0.24**	**81**	**7**	-
COLMAP	4.99	0.12	66	12	0/32
Ours (COLMAP init)	**2.48**	0.12	**81**	**0**	-

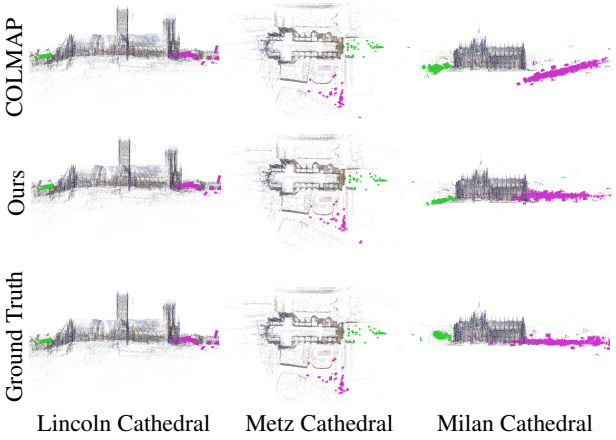

Figure 6. **Grounding Multiple Meta Images**. Above we show three scenes containing two meta-images per scene (visualized in **green** and **purple**), both grounded to the scene's global reference model. We show both the COLMAP initialization, and our final result. Ground truth reconstructions are provided on the bottom.

5.3. Comparison to Feed-Forward 3D Models

We compare our method against DUSt3R [58], MASt3R [27], VGGT [56] and π^3 [60], recent feed-forward reconstruction models. We conduct two sets of experiments to demonstrate that these models struggle in our problem setting: (i) meta-meta reconstructions, which reconstructs two meta-images belonging to the same physical scene without a reference model, and (ii) meta-to-reference reconstructions, which aligns a single meta-image to the reference model. We perform these

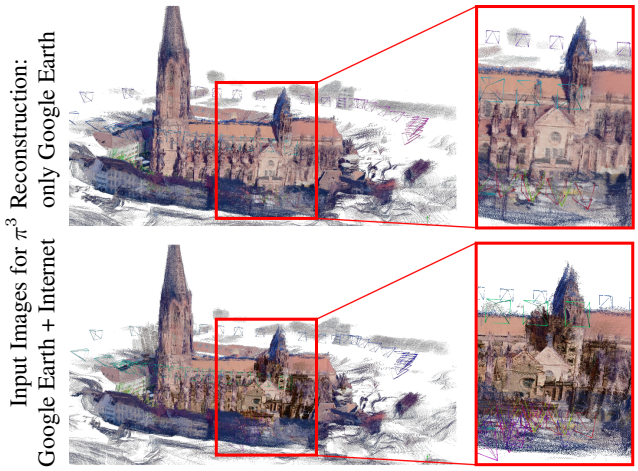

Figure 7. **Aligning a meta-image to a reference model with π^3.** As illustrated above, while π^3 successfully registers the Google Earth images, it struggles to correctly align the Internet images in this model; see, for instance, the **ghost structure** in the center of the red box on the bottom row. The top result, reconstructed from Google Earth images only, is shown for reference.

Table 2. **Comparison with Feed-Forward 3D Models.** We report geodesic rotation errors over two settings, comparing two meta-images directly ($\Delta R_{\mathcal{I}\leftrightarrow\mathcal{I}}$) and comparing a single meta-image with a reference model ($\Delta R_{\mathcal{I}\leftrightarrow\mathcal{M}}$). As illustrated below, our method significantly outperforms all recent feed-forward 3D models by an order of magnitude in both settings.

Methods	$\Delta R^{\circ}_{\mathcal{I}\leftrightarrow\mathcal{M}} \downarrow$	$\Delta R^{\circ}_{\mathcal{I}\leftrightarrow\mathcal{I}} \downarrow$
DUSt3R	54.40	29.27
MASt3R	24.18	12.52
VGGT	51.69	24.63
π^3	68.46	45.80
Ours (COLMAP Init)	**2.59**	**1.48**

experiments over the seven scenes in *WikiEarth* that contain multiple meta-images; this yields 16 and 32 meta-meta and meta-to-reference comparisons, respectively.

As these methods cannot handle hundreds of images, we subsample 45 images each time (larger collections yield OOM on our A5000 GPU). For the alignment with the reference model, we subsample 35 from the reference model and 10 from each meta-image. As π^3 is more memory efficient, for this baseline we subsample 180 images; for the alignment with the reference model we subsample 150 and 30 images from the reference model and meta-image respectively. We sample evenly from each meta-image for the experiment without the reference model. Each run is repeated five times. To quantify performance, we measure geodesic rotation error [2] between image pairs: $\Delta R_{\mathcal{I}\leftrightarrow\mathcal{I}}$ denotes the average error for meta-meta reconstructions and $\Delta R_{\mathcal{I}\leftrightarrow\mathcal{M}}$ denotes the average error for meta-to-reference alignment.

Results are reported in Table 2. As shown, our method outperforms all baselines by an order of magnitude across both metrics, demonstrating that these models cannot cope with our challenging problem setting. Figure 7 illustrates a typical failure mode of meta-to-reference alignment. Additional qualitative results are provided in the supplementary.

5.4. Generalization to Different Reference Model

To demonstrate the generalization of our method to global reference models obtained from various sources, we created global reference models from drone videos (downloaded from YouTube). We constructed the reference models automatically by sampling the frames and recovering the trajectory using SfM. Results are presented in Figure 8.

5.5. Ablation Study

To assess the contributions of different components in our method, we conduct ablation studies on both the semantic features and robust optimization techniques. These studies help isolate the effects of key optimizations and demonstrate the importance of each component. We conduct these ablations using the COLMAP initialization. The results of these ablations are summarized in Table 3. Additional ablations are provided in the supplementary.

Semantic Features Ablations: We explore the impact of different photometric and feature-based optimization methods by replacing our primary semantic features with alternative approaches. This includes a photometric loss and other semantic feature extractors. Photometric loss, commonly used in inverse optimization methods like iNeRF [64] and VF-NeRF [46], effectively aligns scenes through differentiable rendering in controlled environments. However, this method struggles on our "in-the-wild" dataset, as our results show it significantly underperforms, with high rotation error (ΔR of 6.48) and translation error (ΔT of 0.38), likely due to large color variations between meta-images and the low-quality reference model.

We also evaluate LSeg [28], which has been applied successfully in semantic segmentation tasks. While potentially beneficial in scenes with clear semantic structures, LSeg fails to outperform even the initialization (COLMAP), yielding a translation error of 0.34. This suggests that LSeg, though effective for segmentation, lacks the robustness needed for large-scale scene alignment in this challenging setting. Finally, we experiment with DINOv2 + DVT [62], a variant of DINO designed to reduce grid-like artifacts in feature maps. While DINOv2 + DVT performs better than photometric loss and provides reasonable alignment results, we find that it falls short of the performance achieved using standard DINOv2 features, and therefore we did not utilize these semantic features in our pipeline.

Robust Optimization Ablations: To evaluate the effect of our robust optimization scheme, we compare it to several

COLMAP Ours

Figure 8. **Generalization to Drone-Based Reference Models.** We evaluate our method using a reference model reconstructed from drone video frames sourced from YouTube. As illustrated above, our approach significantly improves the alignment over the COLMAP baseline, which serves as our initialization.

Table 3. **Ablations**. We provide several experiments ablating our semantic features (middle) and robust optimization scheme (bottom). All the ablations where initialized with the COLMAP baseline. Best results are marked in bold. As illustrated below, the full method consistently outperforms all alternatives.

Methods	$\Delta R^\circ \downarrow$	$\Delta T \downarrow$	MTA%$\uparrow$	$O\% \downarrow$
Ours (COLMAP init)	**2.48**	**0.12**	**81**	**0**
Semantic Features Ablations				
Photometric Optimization	6.48	0.38	72	22
LSeg [28]	4.78	0.34	62	19
DINOv2 + DVT [62]	2.86	0.14	78	**0**
Robust Optimization Ablations				
w/o LTS	3.78	0.19	69	3
Fixed LTS	2.78	0.14	72	**0**
IRLS [10]	3.51	0.18	72	3
L2	4.21	0.20	75	3

baselines: (i) removing the robust optimization technique (w/o LTS), (ii) fixing the set of ignored images according to the first iteration in LTS (fixed LTS) (iii) replacing it with an alternative robust optimization method (IRLS [10]), and (iv) using $L2$ loss instead of $L1$. The ablation results are presented in the lower part of Table 3.

Both the IRLS and w/o LTS ablations show higher translation errors than the COLMAP baseline, highlighting the importance of LTS in optimizing alignment. Notably the ablation methods show up to 3% outliers, indicating that most scenes converge, though not as effectively as with the full method.

The fixed LTS ablation also outperforms COLMAP but is slightly surpassed by the full approach. This indicates that a soft selection of images for optimization is more effective than a fixed cutoff, adjusting the ignored image set based on the current loss landscape.

We further analyze this phenomenon in the supplementary. In particular, we show histograms depicting the number of times images are ignored throughout the optimization procedure. Our analysis reveals that a majority of the images are consistently ignored (or not), while the relative loss of some images perturbs across iterations, further illustrating that our robust optimization scheme indeed yields a soft selection mechanism allowing for achieving stable convergence with improved performance.

Our ablation studies highlight the importance of both semantic features and robust optimization techniques in achieving superior alignment. The results confirm that our full method consistently outperforms the ablation alternatives, with key optimizations such as LTS and semantic features based approaches playing critical roles in improving accuracy and convergence stability.

6. Conclusion

In this work, we proposed an approach for grounding partial 3D reconstructions *in the wild* to a reference Gaussian Splatting model. Technically, this amounts to aligning in-the-wild images to a model obtained from pseudo-synthetic renderings. To solve it, we frame grounding as an inverse optimization problem and introduce a semantic feature-based robust optimization solution that is capable of handling outliers. Additionally we created *WikiEarth*, a new benchmark dataset that contains 3D reconstructions of landmarks associated internet photo collections previously assembled in the Wikiscenes dataset, registered with reference models obtained via images rendered from Google Earth Studio.

As with other optimization-based approaches, our method remains sensitive to initialization, particularly when the initial alignment is far from the correct solution. Performance also becomes less reliable for very small image collections, which fall outside the regime our framework is designed to address. A more detailed analysis of these limitations is provided in the supplementary material. Future works can leverage the non overlapping aligned meta-images in *WikiEarth* to train sparse reconstruction pipelines. Another promising direction is to combine the global coverage of pseudo-synthetic imagery with the rich appearance details of Internet photographs to construct stronger hybrid 3D scene representations. Finally, incorporating language-based semantic features could unlock downstream applications such as language-guided scene grounding and navigation across large-scale environments.

Acknowledgments This work was partially supported by the Israel Science Foundation grants 2510/23 and 2132/23.

References

[1] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 3

[2] Hana Bezalel, Dotan Ankri, Ruojin Cai, and Hadar Averbuch-Elor. Extreme rotation estimation in the wild, 2025. 2, 7

[3] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes, 2021. 2

[4] Jiahao Chang, Yinglin Xu, Yihao Li, Yuantao Chen, and Xiaoguang Han. Gaussreg: Fast 3d registration with gaussian splatting, 2024. 3

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 3

[6] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning, 2021. 2

[7] Yu Chen and Gim Hee Lee. Dreg-nerf: Deep registration for neural radiance fields, 2023. 3

[8] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration, 2020. 3

[9] Andrea Cohen, Johannes L Schönberger, Pablo Speciale, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. Indoor-outdoor 3d reconstruction alignment. In *European Conference on Computer Vision*, pages 285–300. Springer, 2016. 2

[10] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Gunturk. Iteratively re-weighted least squares minimization for sparse recovery, 2008. 8

[11] Shay Dekel, Yosi Keller, and Martin Cadik. Estimating extreme 3d image rotations using cascaded attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2588–2598, 2024. 2

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description, 2018. 4, 5

[13] Bertram Drost and Slobodan Ilic. 3d object detection and localization using multimodal point pair features. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 9–16. IEEE, 2012. 3

[14] Chen Dudai, Morris Alper, Hana Bezalel, Rana Hanocka, Itai Lang, and Hadar Averbuch-Elor. Halo-nerf: Learning geometry-guided semantics for exploring unconstrained photo collections. In *Computer Graphics Forum*, page e15006. Wiley Online Library, 2024. 4

[15] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, Hanwen Jiang, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object, in any scene, with one reference. *arXiv preprint arXiv:2305.15727*, 2023. 2

[16] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi. nerf2nerf: Pairwise registration of neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9354–9361, 2023. 3

[17] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan. 3d object recognition in cluttered scenes with local surface features: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(11): 2270–2287, 2014. 3

[18] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, Jianwei Wan, and Ngai Kwok. A comprehensive performance evaluation of 3d local feature descriptors. *International Journal of Computer Vision*, 116, 2015. 3

[19] Gerd Häusler and D Ritter. Feature-based object recognition and localization in 3d-space, using a single video image. *Computer Vision and Image Understanding*, 73(1):64–81, 1999. 3

[20] Itan Hezroni, Amnon Drory, Raja Giryes, and Shai Avidan. Deepbbs: Deep best buddies for point cloud registration, 2021. 3

[21] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. Nerf-rpn: A general framework for object detection in nerfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23528–23538, 2023. 3

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 4

[24] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields, 2023. 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3

[26] Shai Krakovsky, Gal Fiebelman, Sagie Benaim, and Hadar Averbuch-Elor. Lang3d-xl: Language embedded 3d gaussians for large-scale scenes. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025. 4

[27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 1, 2, 3, 5, 6

[28] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 3, 7, 8

[29] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 5

[30] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 2

[31] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement, 2021. 4

[32] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed, 2023. 4, 5

[33] Jianlin Liu, Qiang Nie, Yong Liu, and Chengjie Wang. Nerf-loc: Visual localization with conditional neural radiance field. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9385–9392. IEEE, 2023. 3

[34] Ricardo Martin-Brualla, Yanling He, Bryan C Russell, and Steven M Seitz. The 3d jigsaw puzzle: Mapping large indoor spaces. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014. 2

[35] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[36] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022. 3

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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 Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 4

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

[39] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Visual localization using imperfect 3d models from the internet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13175–13186, 2023. 3

[40] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting, 2024. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[42] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388): 871–880, 1984. 4

[43] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the feature: Learning robust camera localization from pixels to pose, 2021. 3

[44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[45] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 4, 5

[46] Leo Segre and Shai Avidan. Vf-nerf: Viewshed fields for rigid nerf registration, 2024. 2, 3, 4, 7

[47] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding, 2023. 3

[48] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. 2006. 1

[49] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 1

[50] Chris Sweeney, Victor Fragoso, Tobias Hollerer, and Matthew Turk. Large scale sfm with the distributed camera model, 2016. 4, 5

[51] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *arXiv preprint arXiv:2406.11819*, 2024. 5

[52] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes, 2021. 3

[53] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21674–21684, 2025. 2

[54] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory, 2024. 3

[55] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2

[56] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 2, 3, 5, 6

[57] Ruibo Wang, Song Zhang, Ping Huang, Donghai Zhang, and Wei Yan. Semantic is enough: Only semantic information for nerf reconstruction. In *2023 IEEE International Conference on Unmanned Systems (ICUS)*, page 906–912. IEEE, 2023. 3

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

[59] Yue Wang and Justin M. Solomon. Deep closest point: Learning representations for point cloud registration, 2019. 3

[60] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning, 2025. 2, 5, 6

[61] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 428–437, 2021. 2, 5

[62] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers, 2024. 7, 8

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

[64] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 2, 3, 4, 6, 7

[65] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022. 2

[66] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3d registration with maximal cliques, 2023. 3

[67] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3

[68] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. 2016. 3

[69] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 3, 4