EMPOWERING FEED-FORWARD RECONSTRUCTION MODELS WITH METRIC SCALE VIA SATELLITE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

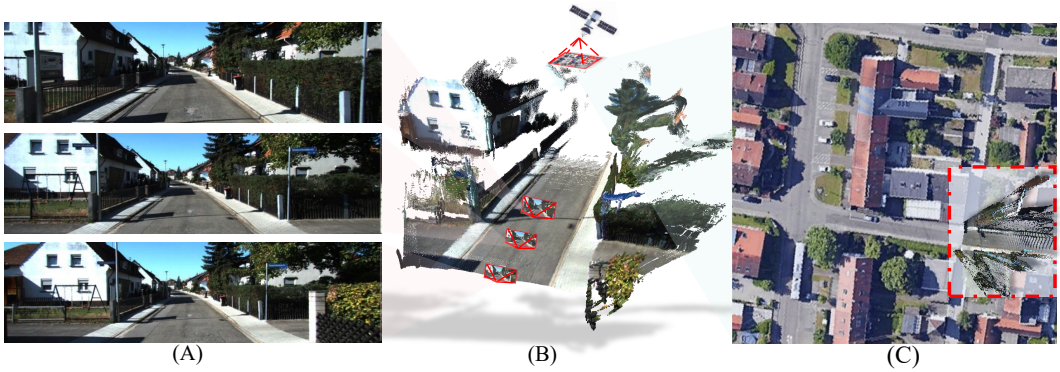
Feed-forward 3D reconstruction models have recently shown strong generalization across diverse scenes, yet most of them recover geometry only up to an unknown global scale. This scale ambiguity limits their use in applications that require metric understanding of the environment. Existing metric reconstruction methods commonly rely on large-scale metric annotations or accurate camera calibration, both of which are costly or unreliable in many real-world settings. We propose a satellite-guided framework for resolving scale ambiguity in outdoor feed-forward 3D reconstruction. The key idea is to use readily available satellite imagery as a global metric reference. Given coarse position and heading estimates, our method retrieves a local satellite patch and integrates it with a feed-forward reconstruction backbone through bidirectional cross-view interaction. By enforcing consistency between the reconstructed scene and the satellite reference, the model infers absolute scale, refines scene geometry, and estimates camera pose in a metric coordinate frame. Experiments on KITTI, nuScenes, and Oxford RobotCar demonstrate improvements in metric depth estimation, multi-view point-cloud reconstruction, and camera localization under the evaluated cross-dataset and geographic transfer protocols.

1 INTRODUCTION

Visual geometry reconstruction is a fundamental problem in computer vision. Reliable reconstruction is critical for downstream applications such as augmented reality (Cao et al., 2020), autonomous driving (Liao et al., 2025), and embodied AI (Duan et al., 2022). Compared with traditional pipelines such as Structure-from-Motion (SfM) (Özyeşil et al., 2017) and Multi-View Stereo (MVS) (Furukawa & Hernández, 2015), recent feed-forward 3D reconstruction methods (Wang et al., 2024; 2025b;d) enable efficient end-to-end inference without iterative optimization and show strong robustness across diverse scenes. However, most feed-forward methods recover only scale-agnostic geometry: the absolute metric scale of the scene remains ambiguous. This limitation restricts their use in real-world tasks that require metric understanding of 3D space.

Recent works (Piccinelli et al., 2025; Wang et al., 2025c; Lin et al., 2025) have recognized the importance of metric depth and attempted to mitigate scale ambiguity by training models on large-scale datasets with metric depth annotations. Through such supervision, models can learn object-size priors and produce approximate metric predictions. While this strategy improves metric estimation in familiar environments, its metric predictions depend on learned priors, which may generalize less effectively to unseen scenes or object categories. Another line of work, represented by feed-forward reconstruction frameworks (Wang et al., 2024; 2025d), focuses on scale-agnostic geometric representations. These methods learn transferable relative geometry, but the lack of absolute scale limits their applicability to metric tasks. Other approaches explore depth-supervised feature learning (Liu et al., 2024) or dual-pixel sensing (Ashida et al., 2024) to recover metric scale. These approaches exploit different sources of metric information, from supervised scene priors to additional sensor measurements. These challenges stem from the inherent scale ambiguity of visual perception: image observations alone do not uniquely determine absolute scene scale, and multiple 3D reconstructions differing by a global scale factor can explain the same visual evidence. Recovering metric geometry therefore requires external references or additional geometric constraints.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(A) (B) (C)

Figure 1: Using coarse position and heading estimates together with retrieved satellite patches, our method resolves the scale ambiguity of foundation-model-based 3D estimation and achieves metric scene reconstruction together with absolute camera pose estimation. In (C), the red box illustrates the projection of the reconstructed metric scene onto the satellite image using the estimated absolute camera pose, demonstrating accurate geometric alignment.

We propose to introduce metric scale information into feed-forward 3D reconstruction by leveraging satellite imagery as an external reference. Satellite images are widely available and encode the physical layout of outdoor environments in a metric coordinate system. A geographically referenced patch supplies scene-specific spatial information that complements the geometric priors of a pretrained reconstruction model. Our objective is to integrate this reference into joint scale recovery, camera localization, and geometric refinement. They therefore provide complementary cues for resolving scale ambiguity without requiring accurate camera calibration at inference time, while training uses LiDAR-derived metric depth and camera-pose supervision.

Using satellite imagery for geometric reasoning requires aligning ground-view observations with the corresponding satellite region. However, precise GPS measurements are not always available or reliable in real-world scenarios (Xia et al., 2023). We use a coarse GPS estimate to retrieve a local satellite patch covering the scene, together with a coarse heading estimate for initialization. From this coarse initialization, our framework jointly aligns ground-view observations with the satellite patch to infer absolute camera pose and metric scene geometry. This design allows the model to exploit satellite imagery as a global spatial reference and enables unified estimation of camera localization and metric 3D reconstruction.

The central challenge is to couple scale recovery with cross-view alignment: the reconstruction is initially expressed in arbitrary units, while its placement in the satellite map is uncertain. We address this coupling with bidirectional cross-view interaction and a scene-to-satellite alignment objective. Satellite features aggregate ground-view geometry to estimate global scale and coordinate alignment; reverse interaction supplies satellite context for per-point geometric refinement. This decomposition connects a shared metric reference to both camera pose and dense scene geometry.

We evaluate this formulation through complementary reconstruction and localization tasks. Cross-dataset depth and point-cloud experiments assess transfer of the satellite-assisted system, while scale-aligned evaluations examine geometry after removing global scale error. Cumulative ablations and extensions to two additional backbones assess the roles of geometric correction and the applicability of the augmentation beyond the main backbone.

In summary, our contributions are as follows:

- We formulate satellite-assisted metric reconstruction from coarse camera poses, combining a georeferenced scene reference with pretrained feed-forward geometry.
- We develop bidirectional cross-view interaction and scene-to-satellite alignment for joint scale recovery, absolute camera pose estimation, and dense geometric refinement.
- We demonstrate the framework on metric depth, point-cloud reconstruction, and camera localization, with cross-dataset evaluation and extensions to multiple reconstruction backbones.

2 Related Work

2.1 Feed-forward 3D reconstruction

Feed-forward 3D reconstruction has recently emerged as an efficient alternative to traditional pipelines based on structure-from-motion (Schönberger et al., 2016; Iglhaut et al., 2019) and multi-view stereo (Hartley & Zisserman, 2003; Goesele et al., 2006). These methods infer 3D structure directly from images through a single neural-network forward pass. DUSt3R (Wang et al., 2024) and its successors (Leroy et al., 2024a; Smart et al., 2024; Wang & Agapito, 2025) formulate geometric 3D reasoning as a feed-forward visual prediction problem, enabling direct reconstruction from image pairs without conventional optimization. Fast3R (Yang et al., 2025) extends feed-forward reconstruction to large-scale settings and reconstructs scenes from more than one thousand images in a single forward pass. Methods such as FLARE (Zhang et al., 2025) further integrate geometry, appearance, and camera estimation into a unified feed-forward framework under uncalibrated and sparse-view conditions. More recently, VGGT (Wang et al., 2025b;a) and π^3 (Wang et al., 2025d) introduce visual geometric priors within Transformer-based architectures, enabling global reasoning over multi-view inputs. These methods show impressive generalization across scenes and datasets. Many feed-forward approaches recover relative geometry, while metric depth models (Lin et al., 2025) learn metric information through large-scale training. AMB3R (Wang & Agapito, 2026) further enables metric-scale feed-forward reconstruction through a sparse volumetric backend for geometric reasoning. Our approach augments pretrained reconstruction with a georeferenced satellite reference, jointly estimating its metric scale and placement in that reference. This formulation connects metric reconstruction to cross-view localization.

2.2 Satellite-image-based camera pose estimation

Satellite-image-based camera pose estimation localizes a ground camera by comparing ground-level observations with satellite imagery. Retrieval-based methods (Zhu et al., 2021; Vivanco Cepeda et al., 2023; Vyas et al., 2022; Shi et al., 2022; 2020) identify the most similar satellite patches from an image database, but their accuracy and efficiency are often constrained by database size. Other methods refine an initial GPS estimate. One line of work (Shi & Li, 2022; Shi et al., 2023) iteratively optimizes the alignment between ground and satellite images, while another line (Xia et al., 2023; Lentsch et al., 2023; Tong et al., 2025) samples candidate poses from satellite imagery and selects the best one by evaluating ground-image consistency. A third class of methods (Song et al., 2023; Wang et al., 2023; Xia & Alahi, 2025) establishes cross-view correspondences and estimates relative pose from matched points. Beyond camera localization, Cross-View Splatter (Turkulainen et al., 2026) combines georeferenced ground images and orthorectified satellite imagery to predict Gaussian representations for novel-view synthesis. These approaches demonstrate that satellite imagery provides a low-cost and widely available geographic reference. Our focus is joint metric scale recovery and coarse-pose refinement through scene-to-satellite alignment, with dense point maps and absolute camera poses as outputs.

3 Method

3.1 Problem Formulation

Preliminary. Recent feed-forward 3D reconstruction methods have demonstrated strong performance across diverse datasets and scenarios. Given a single image or a sequence of N images G_N, a base feed-forward model $\phi(\cdot)$ predicts camera motion and scene geometry as

$$(T_N, X_N) = \phi(G_N), \tag{1}$$

where $T_N \in \mathbb{R}^{N\times 4\times 4}$ denotes the estimated camera poses, and $X_N \in \mathbb{R}^{N\times H_G\times W_G\times 3}$ represents the reconstructed dense 3D point maps, where H_G and W_G denote the spatial dimensions of each point map. The pose of the i-th camera is written as $T_i = [T_i^r, T_i^t]$, where $T_i^r \in \mathrm{SO}(3)$ and $T_i^t \in \mathbb{R}^3$ denote the camera-to-world rotation and camera center in the reconstruction frame, respectively.

For the scale-agnostic backbones considered here, both camera motion and scene geometry are defined only up to an arbitrary global scale. While such representations generalize well, the absence of absolute scale limits their applicability in tasks requiring metric reasoning.

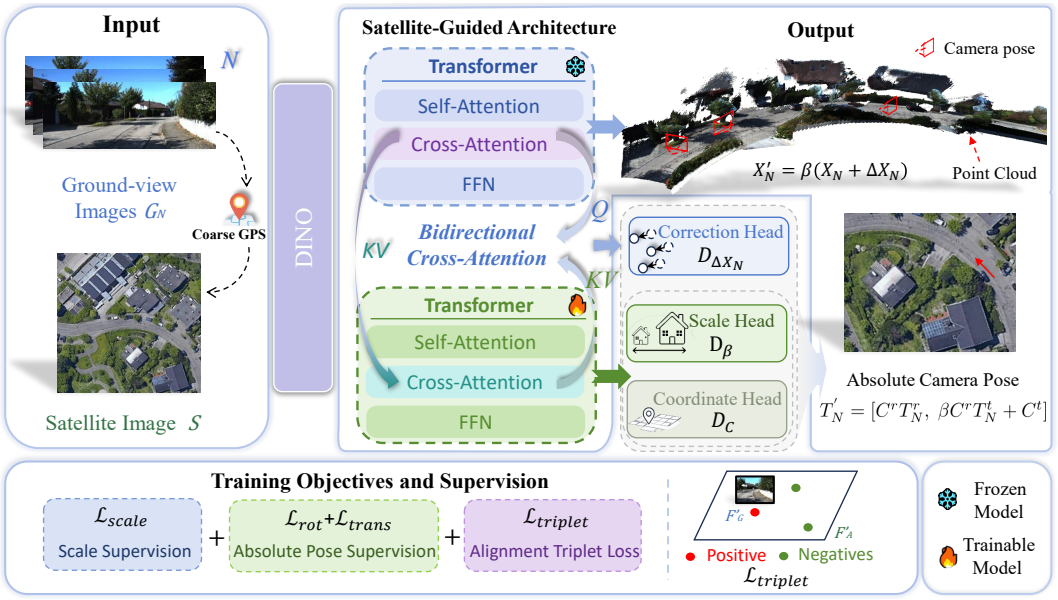


Figure 2: Overview of the proposed framework. Given a coarse ground-view pose, our method retrieves a corresponding satellite patch as a global metric reference. Through bidirectional interaction between satellite and ground-view branches, the network estimates a global scale factor β, a coordinate offset C, and a geometric correction term ΔX_N, enabling metric scene reconstruction and absolute camera pose estimation in the satellite coordinate frame.

Our approach. Our goal is to recover metric 3D geometry and accurate vehicle localization given coarse GPS position and heading estimates. We retrieve satellite imagery according to the coarse GPS measurements and use it as an external metric reference. The patch's ground sampling distance defines the conversion between satellite pixels and horizontal metric distances. Given the scale-agnostic output of the base model, our method predicts a scale factor β, a coordinate offset C, and a geometric correction ΔX_N:

$$T'_N = [C^r T^r_N,\ \beta C^r T^t_N + C^t], \quad X'_N = C^r[\beta(X_N + \Delta X_N)] + C^t, \tag{2}$$

where β denotes the global scale factor, X_N and ΔX_N are expressed in the original reconstruction frame, and $C = (C^r, C^t)$ maps the scaled geometry to the satellite frame.

3.2 Satellite-Augmented Metric Scale Estimation

We adopt π^3 as a representative backbone, although the design generalizes to other feed-forward reconstruction models. We freeze the pretrained base model ϕ and clone it into a trainable satellite branch: the frozen branch processes ground-view images and retains its pretrained parameters, while the satellite branch processes satellite imagery and learns to inject metric cues.

Both branches share the frozen encoder. DPT fusion (Ranftl et al., 2021), RGB-guided joint bilateral upsampling (Fu et al., 2024), and a shared projection produce dense matching features f_G, f_S.

Since satellite and ground-view images differ drastically in viewpoint, time of capture, and appearance, we design an asymmetric cross-attention mechanism operating at different depths of the decoder to establish reliable cross-view correspondence. Let $F_S \in \mathbb{R}^{L_S \times d}$ and $F_G \in \mathbb{R}^{L_G \times d}$ denote satellite and ground-view features, where L_S and L_G count tokens and d is the feature dimension. In the primary direction, satellite features query ground-view features throughout the decoder:

$$\tilde{F}_S = \mathrm{softmax}\left(\frac{Q_S K_G^\top}{\sqrt{d}}\right) V_G, \tag{3}$$

where $Q_S = F_S W_Q^S$, $K_G = F_G W_K^S$, $V_G = F_G W_V^S$. Running this interaction across all decoder layers allows the satellite branch to progressively aggregate ground-view geometry, building a rich cross-view representation essential for accurate scale and coordinate estimation. In the reverse direction,

Table 1: Monocular depth estimation results on the KITTI dataset. Our satellite-guided modules are trained on nuScenes and evaluated on KITTI with the pretrained backbone, achieving the best reported performance.

	↓AbsRel	↓SqRel	↓ RMSE	↓ RMSE_{log}	↑ δ_1	↑ δ_2	↑ δ_3
DepthPro	0.1833	0.7400	3.6684	0.1872	0.7361	0.9813	0.9972
MoGeV2	0.1822	0.6595	3.8175	0.2191	0.6142	0.9786	0.9970
DA3	0.2263	1.0362	4.4589	0.2198	0.5512	0.9860	0.9964
Ours	**0.1371**	**0.5361**	**3.4067**	**0.1558**	**0.8542**	**0.9881**	**0.9976**

applied only in the final decoder layers, ground-view features query satellite features:

$$\tilde{F}_G = \text{softmax}\left(\frac{Q_G K_S^\top}{\sqrt{d}}\right) V_S, \tag{4}$$

where $Q_G = F_G W_Q^G$, $K_S = F_S W_K^G$, $V_S = F_S W_V^G$. This lightweight reverse interaction injects satellite-derived metric and structural cues into the ground-view representation, enabling geometric refinement. Restricting this interaction to the final layers introduces satellite guidance late in the decoder while keeping the pretrained parameters fixed.

Three task-specific heads decode the cross-view features:

$$\beta = D_\beta(\tilde{F}_S), \quad C = D_C(\tilde{F}_S), \quad \Delta X_N = D_{\Delta X_N}(\tilde{F}_G). \tag{5}$$

D_β applies global average pooling and a linear layer to predict the scalar scale β; D_C is an MLP regressing the alignment parameters $C = [C^r, C^t]$; and $D_{\Delta X_N}$ predicts per-point geometric corrections ΔX_N that compensate for reconstruction errors. We enforce positivity of the predicted scale β and obtain $C^r \in \text{SO}(3)$ by applying SVD-based orthogonalization with determinant correction to the regressed 3×3 matrix. For numerical stability, we clip the argument of $\arccos$ to $[-1+\epsilon, 1-\epsilon]$, where $\epsilon = 10^{-6}$.

3.3 Training Objectives and Supervision

Scale factor supervision. Whereas scale-agnostic reconstructions (Wang et al., 2025b;d) require scale alignment for metric evaluation, our method predicts β at inference time via satellite imagery. We obtain a reference scale $\hat{\beta}$ from valid ground-truth depth measurements and the corresponding unscaled depth predictions, and supervise the scale head with an ℓ_2 loss in log-space:

$$\mathcal{L}_{\text{scale}} = \left(\log \beta - \log \hat{\beta}\right)^2. \tag{6}$$

The log-space loss penalizes the ratio between the predicted and reference scales, assigning the same penalty to equal multiplicative errors.

Absolute pose supervision. By aligning the reconstructed scene with satellite imagery, we localize the camera in the satellite reference frame and recover its absolute pose. Following Eq. 2, the absolute camera pose is constructed as $T_N' = [C^r T_N^r,\ \beta C^r T_N^t + C^t]$. Let $\hat{P}_N = [\hat{P}_N^r, \hat{P}_N^t]$ denote the ground-truth absolute camera pose in the satellite coordinate frame. We define rotation and translation losses as

$$\mathcal{L}_{\text{rot}} = \arccos\Big((\text{Tr}\Big((C^r T_N^r)^\top \hat{P}_N^r\Big) - 1)/2\Big), \quad \mathcal{L}_{\text{trans}} = \mathcal{H}(\hat{\beta} C^r T_N^t + C^t - \hat{P}_N^t), \tag{7}$$

where $\mathcal{H}(\cdot)$ denotes the Huber loss. To decouple pose learning from scale prediction during training, we substitute the depth-derived reference scale $\hat{\beta}$ for the predicted scale in the translation loss.

Satellite alignment consistency supervision. To encourage effective use of satellite imagery for metric reasoning, we explicitly supervise alignment consistency between the reconstructed scene and the corresponding satellite image in a bird's-eye-view (BEV) representation. Given the geometry estimated by the foundation model and the recovered scale and coordinate offset, we project the dense matching features f_G into BEV space:

$$f_G^{\text{BEV}} = P(f_G,\ X_N'), \tag{8}$$

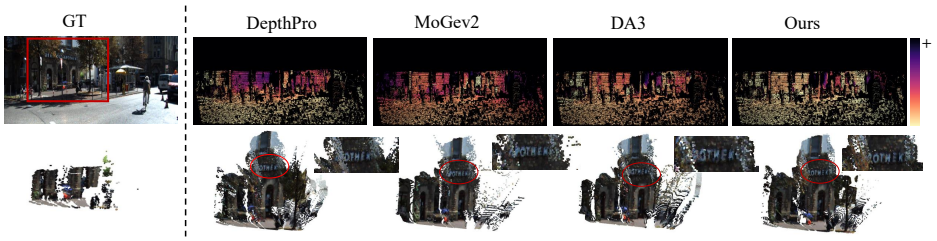


Figure 3: Qualitative comparison of monocular metric depth estimation. The input image highlights the building evaluated below. The top row compares predicted depth errors, and the bottom row shows the corresponding building point clouds for ground truth and each method.

Table 2: Multi-view metric depth estimation on the KITTI dataset. In the upper part, all methods are aligned to the ground-truth metric scale. In the lower part, metric-scale predictions are directly compared against ground truth.

	Scale	↓AbsRel	↓SqRel	↓ RMSE	↓ RMSE_{log}	↑ δ_1	↑ δ_2	↑ δ_3
Fast3R	GT	0.1024	0.5489	3.9426	0.1501	0.9053	0.9789	0.9935
Mast3R		0.0834	0.6435	4.6779	0.3605	0.9445	0.9774	0.9861
PI3		0.0464	0.1488	2.1142	0.0748	0.9828	0.9970	**0.9989**
VGGT		0.0703	0.2904	3.2424	0.1064	0.9618	0.9899	0.9966
DA3		0.0473	0.2601	2.5869	0.0875	0.9754	0.9938	0.9978
Ours		**0.0353**	**0.1219**	**2.0039**	**0.0634**	**0.9896**	**0.9972**	**0.9989**
DA3	Pred	0.2448	1.1920	5.0979	0.3014	0.3588	0.8582	0.9754
Ours		**0.1226**	**0.3605**	**2.9460**	**0.1475**	**0.8590**	**0.9988**	**0.9989**

where $P(\cdot)$ projects features onto the horizontal BEV plane by differentiable bilinear splatting. Contributions to the same cell are accumulated and normalized by their summed weights plus 10^{-6}; cells receiving no contribution remain zero. Confidence and positive-depth soft gates down-weight unreliable points. The DPT–JBU pathway supplies spatially detailed features, so the projected BEV map can be compared with f_S at a resolution sufficient for localization. Here, X'_N is expressed in the satellite frame using the predicted scale and pose transformation in Eq. 2; the ground-truth pose provides supervision rather than replacing this transformation. We supervise this correspondence with a triplet-style loss. For each input ground-view image, the ground-truth camera location $(\hat{u}_i, \hat{v}_i)$ on the satellite map is treated as the positive sample, while other satellite locations (u, v) are treated as negatives:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{HW - 1} \sum_{(u,v) \neq (\hat{u}_i, \hat{v}_i)} \log(1 + \exp(\alpha \left(\mathcal{C}_i(\hat{u}_i, \hat{v}_i) - \mathcal{C}_i(u, v)\right))) \,, \tag{9}$$

where H and W are the height and width of the matching cost map. The matching cost $\mathcal{C}_i(u, v)$ is computed as two minus twice the normalized cross-correlation between the BEV-projected ground-view features and the satellite features at candidate location (u, v), with lower values indicating better matches. We set $\alpha = 10$ to control the sharpness of the loss. Normalized cross-correlation is computed over the nonzero support of the projected BEV features, with both feature norms evaluated on the same support and their product lower-bounded by 10^{-6}. The ground-truth location is rounded to the nearest grid cell; all remaining candidate locations are treated as negatives, without excluding or down-weighting spatial neighbors.

4 Experiments

Implementation details. We perturb initial positions by $\pm 20\,\text{m}$ and yaw by $\pm 10°$ (Shi & Li, 2022; Xia et al., 2023). With pretrained reconstruction parameters frozen, Stage A trains cross-view features and scale/pose heads for 20 epochs using feature-based pose, feature-alignment, and scale losses. Stage B loads this checkpoint, freezes the scale/pose heads, and trains cross-view features and the correction head for 8 epochs using point-cloud correction and feature-alignment losses. Training samples one to three ground images.

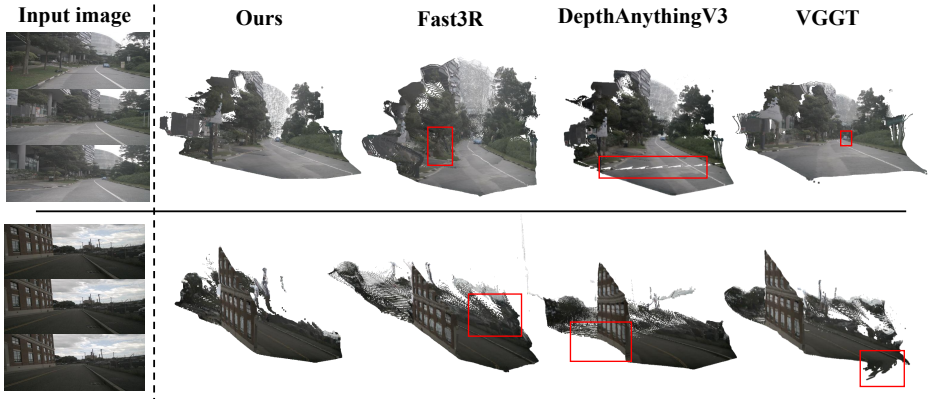


Figure 4: Visualization of multi-view point-cloud estimation. Our method produces reconstructions that are more complete and less noisy than competing approaches.

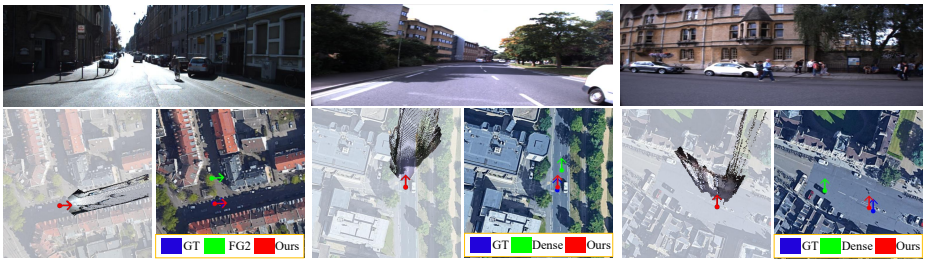
Table 3: Multi-view point-cloud estimation on the nuScenes dataset. Both blocks use rigid registration. The upper block additionally adjusts scale to the reference cloud; the lower block retains the predicted global scale.

	Scale	↓Acc.		↓Comp.		↑N.C.	
		Mean	Med.	Mean	Med.	Mean	Med.
Fast3R	GT	**0.9661**	0.6688	1.8719	0.7425	0.8313	0.9331
Mast3R		1.2950	0.7723	1.4813	0.7723	0.8912	0.9814
PI3		1.1119	0.6518	1.7468	1.0946	**0.8983**	0.9812
VGGT		1.3006	0.8965	1.4231	0.7246	0.8481	0.9612
DA3		1.4775	0.8308	1.7288	0.9614	0.8875	0.9800
Ours		1.0879	**0.5959**	**1.1967**	**0.7086**	0.8956	**0.9886**
DA3	Pred	2.0914	1.2001	2.5409	1.3642	0.8662	0.9748
Ours		**1.4982**	**0.8089**	**1.7042**	**0.9590**	**0.8702**	**0.9792**

4.1 Scene Geometry Estimation

To evaluate scene geometry estimation, we conduct experiments on three tasks: monocular metric depth estimation, multi-view metric depth estimation, and multi-view metric point-cloud estimation. Metric evaluation assesses the recovered scale together with geometry; ground-truth scale alignment evaluates the reconstruction after removing its global scale error. We report both to distinguish scale recovery from changes in reconstruction quality. Cross-dataset evaluation is performed on KITTI (Geiger et al., 2013; Shi & Li, 2022) and nuScenes (Caesar et al., 2019; Fong et al., 2021) to assess generalization. Specifically, we train the satellite-guided modules on KITTI for evaluation on nuScenes, and on nuScenes for evaluation on KITTI, using the pretrained reconstruction backbone. For KITTI, we use the satellite imagery from (Shi & Li, 2022). For nuScenes, we augment the dataset with satellite images to support cross-view supervision. The satellite images are retrieved from Google Maps (Google, 2024) using publicly available GPS annotations, with each source image covering an area of $250\,\mathrm{m} \times 250\,\mathrm{m}$. The nuScenes preprocessing extracts a heading-oriented local crop of approximately $100\,\mathrm{m} \times 100\,\mathrm{m}$ after applying the simulated pose perturbation. Our method targets outdoor scenarios with coarse GPS and heading estimates, using GPS to retrieve satellite imagery. The depth and reconstruction baselines in Tables 1–3 use ground-view images alone. These comparisons evaluate satellite-assisted reconstruction alongside image-only methods under their respective input conditions.

Monocular metric depth estimation. We evaluate monocular depth estimation on KITTI. We train the satellite-guided modules on nuScenes and evaluate the resulting system directly on KITTI, keeping the pretrained backbone fixed. As shown in Table 1, we report Absolute Relative Error (AbsRel), Squared Relative Error (SqRel), Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error ($\mathrm{RMSE}_{\log}$), and the accuracies δ_1, δ_2, and δ_3 at depth-ratio thresholds 1.25, 1.25^2, and 1.25^3, respectively. Compared with existing monocular metric depth estimation methods (Bochkovskii et al., 2024; Wang et al., 2025c; Lin et al., 2025), our approach achieves superior performance



(A) KITTI (B) OxFord RobotCar

Figure 5: Visualization of localization results on the KITTI and Oxford RobotCar datasets. In each group, the top image shows the ground-view image; the bottom-left image projects the reconstructed ground scene onto the satellite image; and the bottom-right image indicates the estimated camera position within the satellite map.

Table 4: KITTI location and orientation error. Best in bold, second-best underlined.

	Same Area				Cross Area			
	↓Loc. (m)		↓Orien. (°)		↓Loc. (m)		↓Orien. (°)	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
CCVPE	1.22	0.62	0.67	0.54	9.16	3.33	1.55	0.84
DenseFlow	1.48	**0.47**	0.49	0.30	7.97	3.52	2.17	1.21
FG2	0.75	0.52	1.28	0.74	7.45	4.03	3.33	1.88
BEVSplat	2.87	2.06	**0.33**	**0.28**	6.20	2.51	**0.33**	**0.28**
Ours	**0.73**	0.49	0.43	0.38	**5.75**	**2.34**	0.52	0.37

Table 5: Location error (m) on Oxford RobotCar. Ground-image orientation is assumed known, as in prior work.

	Test1		Test2		Test3		Overall	
	↓Mean	↓Med.	↓Mean	↓Med.	↓Mean	↓Med.	↓Mean	↓Med.
MCC	1.42	1.10	1.95	1.33	1.94	1.29	1.77	1.24
CVR	1.88	1.47	2.64	1.99	2.35	1.71	2.29	1.72
DenseFlow	1.17	**0.72**	1.76	0.97	1.79	0.92	1.57	0.87
Boosting	2.40	0.91	3.10	1.13	2.86	1.05	2.79	1.02
Ours	**1.02**	0.87	**1.09**	**0.90**	**1.08**	**0.74**	**1.07**	**0.84**

across most metrics. These results support the utility of satellite-assisted metric reconstruction in this cross-dataset setting. Figure 3 provides a qualitative comparison of depth errors and reconstructed building geometry.

Multi-view metric depth estimation. We further evaluate multi-view metric depth estimation on KITTI. The satellite-guided modules are trained on nuScenes, and the resulting system is evaluated directly on KITTI. As shown in the lower part of Table 2, we compare with DepthAnythingV3 (Lin et al., 2025), a strong metric multi-frame depth estimation method. Both methods directly produce metric-scale depth maps that are compared against ground truth. Our approach achieves lower errors and higher threshold accuracies in this satellite-assisted setting. In the upper part of Table 2, we additionally compare with scale-agnostic approaches (Yang et al., 2025; Leroy et al., 2024b; Wang et al., 2025b;d; Lin et al., 2025). All methods in this upper block, including ours, are evaluated after ground-truth scale alignment. Our method also improves the scale-aligned depth metrics, indicating differences in reconstruction quality beyond the predicted global scale.

Multi-view point-cloud estimation. We further evaluate multi-view point-cloud estimation on nuScenes using the satellite-guided modules trained on KITTI and the pretrained reconstruction backbone. We report Accuracy (Acc.), Completion (Comp.), and Normal Consistency (N.C.) in Table 3. Acc. and Comp. measure nearest-neighbor distances from predicted points to reference points and in the reverse direction, respectively. N.C. averages absolute normal agreement over the two directions. Both point-cloud protocols use rotation and translation from a similarity fit followed by rigid ICP refinement; the scale-aligned protocol additionally applies the fitted scale, whereas the metric protocol sets this scale to one. Mean and median statistics are computed within each cloud and then averaged across evaluated clouds. Thus, these metrics assess registered reconstruction quality; absolute localization is evaluated separately. In the lower part of the table, both our method and DepthAnythingV3 (Lin et al., 2025) predict metric geometry and are evaluated against the reference point cloud while retaining their predicted global scale. In the upper part, we compare with scale-agnostic approaches (Yang et al., 2025; Leroy et al., 2024b; Wang et al., 2025b;d; Lin et al., 2025), with ground-truth scale alignment applied to every method in this block, including ours. Our method achieves the best results on most reported metrics under this satellite-assisted setting. As shown in Figure 4, our reconstructions are more accurate and complete, producing denser point clouds with fewer outliers.

Table 6: Ablation of core components (KITTI→nuScenes). Each row adds one module.

	Scale L1↓	Metric↓		Align↓	
		Acc	Comp	Acc	Comp
Base model ϕ	18.742	1.706	17.945	1.112	1.747
+ Sat matching	1.948	1.699	1.842	1.186	1.345
+ Feat align.	1.714	1.635	1.748	1.317	1.500
+ ΔX_N	**1.690**	**1.498**	**1.704**	**1.088**	**1.197**

Table 7: Tests on multiple feed-forward methods.

Backbone		Metric↓		Align↓	
		Acc	Comp	Acc	Comp
VGGT	vanilla	6.602	25.85	**1.301**	1.423
	+Ours	**1.567**	**2.642**	1.376	**1.420**
VGGT-Ω	vanilla	2.875	21.04	1.149	1.525
	+Ours	**1.912**	**2.087**	**1.103**	**1.230**

4.2 Metric Pose Estimation

Our method estimates absolute camera pose in the world coordinate system by aligning the reconstructed scene with satellite imagery. The most closely related task is cross-view localization (Shi & Li, 2022; Shi et al., 2023; Xia & Alahi, 2025; Song et al., 2023), where satellite images are used to determine the geographic position of ground-level images. We evaluate on Cross-View KITTI (Shi & Li, 2022) and Oxford RobotCar (Xia et al., 2022), testing geographic and temporal transfer, respectively.

For KITTI localization, we simulate noisy GPS initialization by perturbing the ground-truth pose with random offsets of $\pm 20\,\mathrm{m}$ in latitude and longitude and $\pm 10^\circ$ in yaw. We report mean and median Euclidean position errors on both datasets and mean and median yaw errors on KITTI. Across both datasets, our method achieves the lowest mean localization error, while BEVSplat yields lower yaw errors on KITTI. On Oxford RobotCar, evaluation assumes known orientation and therefore assesses position estimation only. The KITTI cross-area results further support geographic localization, with complementary strengths in position and yaw estimation. As shown in Figure 5, projecting the reconstructed ground scene onto the satellite image produces precise geometric alignment, providing qualitative support for metric scale recovery and absolute pose estimation.

4.3 Ablation Study

We compare successively trained configurations by adding modules one at a time, training on KITTI and evaluating multi-view point-map estimation on nuScenes (Table 6). Scale L1 measures the mean absolute difference between predicted and reference scale factors. The base model ϕ recovers coherent geometry only up to an arbitrary scale, so predictions in the metric frame remain unreliable. Satellite matching supplies the missing metric reference and accounts for most of the scale recovery, bringing metric completion close to the scale-aligned result. Feature alignment then tightens satellite–ground correspondence and further improves accuracy and completion in the metric frame, although scale-aligned errors increase slightly. Adding ΔX_N improves both metric and scale-aligned reconstruction, supporting geometric refinement beyond scale calibration. This cumulative ablation jointly assesses satellite conditioning and metric-supervised adaptation; their individual contributions are not isolated. Table 7 applies our satellite-guided modules to VGGT (Wang et al., 2025b) and VGGT-Ω (Wang et al., 2026) under the KITTI→nuScenes protocol. Satellite guidance improves metric accuracy and completion on both backbones. On VGGT, scale-aligned completion remains comparable while accuracy error increases; both metrics improve on VGGT-Ω. These results demonstrate metric-scale improvements across the evaluated backbones, although improvements in scale-aligned geometry are not uniform.

5 Conclusion

We augment pretrained outdoor reconstruction with georeferenced satellite imagery to jointly estimate metric scale, absolute camera pose, and geometric correction. Training uses metric supervision; inference uses satellite imagery and coarse pose estimates without accurate ground-camera calibration. Cross-dataset reconstruction, localization, and multiple-backbone evaluations support this formulation, with correction improving both metric and scale-aligned errors in the cumulative ablation. Robustness to larger initialization errors, map misregistration, and satellite–ground temporal mismatch remains unestablished.

AI USE STATEMENT

We used generative AI tools to assist with language editing, LaTeX formatting, and pre-submission feedback for this manuscript. All research ideas, technical content, experimental results, and scientific claims were produced and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts.

REFERENCES

Kohei Ashida, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Resolving scale ambiguity in multi-view 3d reconstruction using dual-pixel sensors. In *European Conference on Computer Vision*, pp. 162–178. Springer, 2024.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

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. *arXiv preprint arXiv:1903.11027*, 2019.

Mingwei Cao, Liping Zheng, Wei Jia, Huimin Lu, and Xiaoping Liu. Accurate 3-d reconstruction under iot environments and its applications to augmented reality. *IEEE Transactions on Industrial Informatics*, 17(3):2090–2100, 2020.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *arXiv preprint arXiv:2109.03805*, 2021.

Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. FeatUp: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024.

Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 2402–2409. IEEE, 2006.

Google. Maps static API. `https://developers.google.com/maps/documentation/mapsstatic/intro`, 2024.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Jakob Iglhaut, Carlos Cabo, Stefano Puliti, Livia Piermattei, James OConnor, and Jacqueline Rosette. Structure from motion photogrammetry in forestry: A review. *Current Forestry Reports*, 5(3): 155–168, 2019.

Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17225–17234, 2023.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024a.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024b.

Liewen Liao, Weihao Yan, Wang Xu, Ming Yang, Songan Zhang, and Hongtei Eric Tseng. Learning-based 3d reconstruction in autonomous driving: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Chunpu Liu, Guanglei Yang, Wangmeng Zuo, and Tianyi Zan. Metricdepth: Enhancing monocular depth estimation with deep metric learning. *arXiv preprint arXiv:2412.20390*, 2024.

Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pp. 501–518. Springer, 2016.

Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17010–17020, 2022.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4064–4072, 2020.

Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. Cvlnet: Cross-view semantic correspondence learning for video-based camera localization. In *Asian Conference on Computer Vision*, pp. 123–141. Springer, 2022.

Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21516–21526, 2023.

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

Zhenbo Song, Jianfeng Lu, Yujiao Shi, et al. Learning dense flow field for highly-accurate cross-view camera localization. *Advances in Neural Information Processing Systems*, 36:70612–70625, 2023.

Shaowen Tong, Zimin Xia, Alexandre Alahi, Xuming He, and Yujiao Shi. Geodistill: Geometry-guided self-distillation for weakly supervised cross-view localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25357–25366, 2025.

Matias Turkulainen, Akshay Krishnan, Filippo Aleotti, Mohamed Sayed, Guillermo Garcia-Hernando, Juho Kannala, Arno Solin, Gabriel Brostow, and Daniyar Turmukhambetov. Cross-View Splatter: Feed-forward view synthesis with georeferenced images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In *European Conference on Computer Vision*, pp. 440–456. Springer, 2022.

Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster vggt with block-sparse global attention. *arXiv preprint arXiv:2509.07120*, 2025a.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14612–14625, June 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, 36:5301–5319, 2023.

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. *arXiv preprint arXiv:2507.13347*, 2025d.

Zimin Xia and Alexandre Alahi. FG^2: Fine-grained cross-view localization by fine-grained feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6362–6372, 2025.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pp. 90–106. Springer, 2022.

Zimin Xia, Olaf Booij, and Julian FP Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3813–3831, 2023.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21936–21947, 2025.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021.