CasaMaestro: Multi-View Panoramas for House-Scale 3D Reconstruction

Yuzhou Ji*, Xiaotian Yang*, and Zhipeng Zhang†

School of Artificial Intelligence, Shanghai Jiao Tong University
https://george-attano.github.io/CasaMaestro
jiyuzhou@sjtu.edu.cn

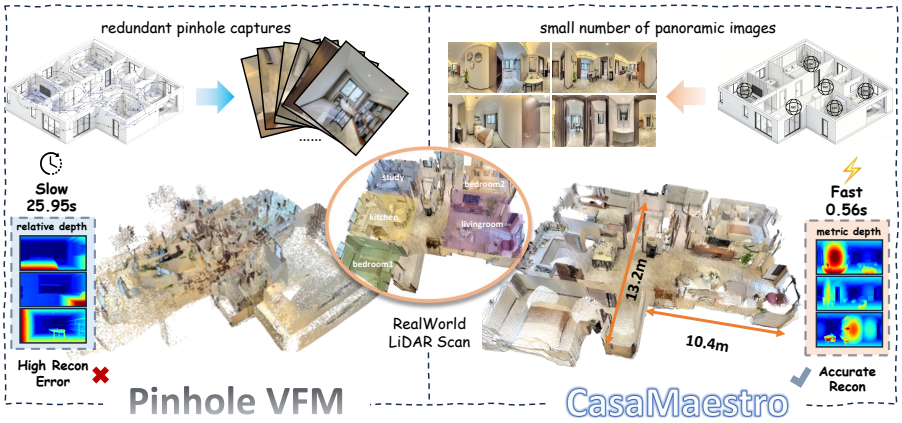


Fig. 1: Demonstration of CasaMaestro. The existing pinhole vision foundation models (VFM) requires dense video streams for scene-level reconstruction, which is consuming in both human effort and computing resources, while also facing severe errors in overly long and short input sequences. We propose CasaMaestro, which takes only sparse indoor panoramas and directly completes house-scale (with multiple rooms) metric reconstruction, providing efficiency along with accuracy.

Abstract. The rise of home-deployed embodied AI systems is driving a growing need for fast, metric 3D reconstruction of residential spaces to support navigation, interaction, and long-horizon task execution. However, the commonly used pinhole-camera 3D reconstruction pipelines struggle to model large indoor residences efficiently due to their limited field of view, to which achieving full coverage across multiple rooms often requires thousands of images and incurs drift from long chains of incremental alignment. In this work, we present CasaMaestro (*Spanish* words meaning "house" and "master"), a feedforward model that can take only twenty to fifty sparse multi-view indoor panoramas as input and

* Equal contribution. † Corresponding author.

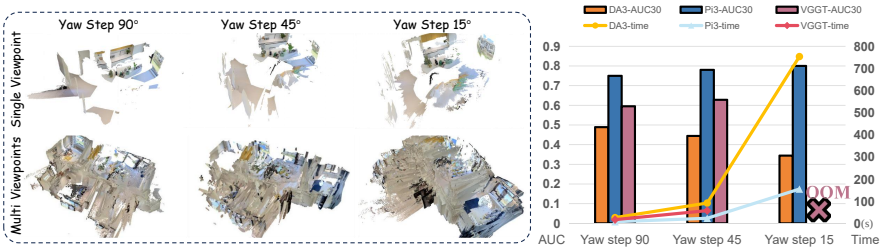


Fig. 2: Illustration of existing problems. Left visualization shows pinhole models either face limited FoV in sparse capture or accumulative error in dense sequence. Right figure shows pose accuracy and processing time under different input density.

directly predicts metric depth along with camera poses, allowing immediate point-cloud reconstruction of the entire house with full coverage. CasaMaestro is the first model that supports house-scale reconstruction with multi-view panoramas. Experiments show that CasaMaestro can robustly provide high quality results in both real-world and synthetic scenes, which can serve as a strong foundation for acquiring house-scale 3D indoor assets to be applied in close-loop simulation.

Keywords: Multi-view panoramas · Point cloud reconstruction · Metric scale reconstruction · Camera pose estimation

1 Introduction

With embodied AI shifting toward in-the-wild operation in everyday homes, the construction of realistic and metrically consistent closed-loop simulation environments becomes increasingly important. In order to bridge the sim-to-real gap, reconstructing virtual 3D residential scenes from real-world capture has served as a crucial role for both training and testing, which, however, still faces many difficulties in practice.

Although LiDAR scanning remains a reliable solution for acquiring high-quality 3D assets of homes, it is costly and hardware-dependent. In contrast, recent 3D vision foundation models [11, 15, 37, 41] suggest a compelling alternative by reconstructing 3D from unconstrained image sequences, enabling 3D lifting for existing and even synthetic imagery without specialized sensors. However, today's large-scale multi-view methods predominantly assume pinhole cameras, whose narrow field of view makes residential capture inefficient and fragile, where dense imaging is required for multi-room coverage and long trajectories exacerbate drift and error accumulation. As shown in Fig. 2, pinhole methods suffer from severe displacement with sparse capture viewpoints that can not be solved (and may even worsen) by increasing single viewpoint scanning density due to accumulative error, while the time and GPU memory consumption are already extremely high. This fact substantially prevents the practical deployment of pure-vision reconstruction for real-home data acquisition.

To improve reconstruction efficiency, several researches have explored using panoramic cameras, yet they are mostly restricted to single-view depth prediction [13, 16] or to small-scale, short-range multi-view settings. For example, PanoSplatt3R [27] provides feed-forward 3DGS reconstruction from only panoramas, but can only handle 2 input views. PanoPose [32] uses separate pose network dedicated for relative pose estimation of panorama pairs to progressively predict long sequences, yet requires large view overlap and consequently a long image sequence for house-level scan. Although SPR [46] supports relatively larger view displacement comparing with PanoPose, it still demands video streams and is solely trained on 5 views setting, leaving the challenge of house-scale multi-view panoramic reconstruction unresolved.

In this work, we present CasaMaestro, the first feedforward model that achieves extrinsic-free multi-view panoramic 3D reconstruction with sparse house-scale capture. Embracing a minimalist design philosophy [15], CasaMaestro is simply built upon a DINOv2 [19] backbone that processes multiple views, paired with dedicated heads for depth and pose prediction. To explicitly tackle the severe view displacements inherent in sparse residential captures, we introduce a lightweight panoramic camera pose decoder. By incorporating a second stage attention mechanism across views, this decoder achieves superior pose estimation accuracy. Moreover, recognizing that existing datasets offer a restricted variety of scenes and viewpoints, we propose a novel panoramic data augmentation strategy via ERP (Equirectangular Projection) remapping, which generates abundant pairs of poses and views, significantly boosting model robustness. Ultimately, CasaMaestro takes only raw panoramas as input to directly predict metric depth and camera poses. This elegant pipeline is remarkably simple yet surprisingly effective, which enables the immediate point cloud reconstruction of entire homes, delivering full coverage and exceptional geometric consistency.

Experiments show that our model is superior than existing methods with specifically **84%** and **119%** improvements of the AUC30 metric in real-world and synthetic scenes with balanced rotation and translation error, and also an average **21.98%** decrease of AbsRel on unseen datasets comparing with the best previous results.

In conclusion, we provide the following contribution:

- We present CasaMaestro, the first feedforward model for metric house-scale panoramic reconstruction.
- We design a lightweight camera pose decoder for stronger pose estimation with large displacement.
- We propose a ERP panoramic data augmentation method for more pose-view pairs with existing data.

2 Related Work

2.1 3D Foundation Models

Feed-forward models has advanced significantly in point cloud reconstruction while predicting multiple 3D attributes, and become 3D Foundation Models for

many downstream vision tasks. The early DUSt3R [40] and MASt3R [12] predict a coupled scene representation but require further post-processing. The following works expand them towards more pipelines [8, 9, 18, 20] and support extra input views [3, 36, 39], but with limited quality compared to traditional optimization.

Recently, multi-view models such as VGGT [37] have made excellent progress in 3D prediction, with many valuable efforts in faster reconstruction [28], more accurate localization [7, 31] and longer sequences [6, 43, 44]. Notably, π^3 [41] employs a fully permutation-equivariant architecture and achieves higher robustness, while MapAnything [11] further enables a broad range of 3D vision tasks in a single feed-forward pass, pushing such models to real-world applications. Later, Depth Anything 3 [15] shows that a single plain transformer is sufficient as a backbone, achieving minimal modeling.

Despite strong performance, the existing 3D foundation models are made upon pinhole cameras, leaving the challenge of panoramic 3D yet to be conquered. In this paper, we aim to provide a panoramic 3D foundation model for future applications. While many researches choose to implement special design for panoramas, we also hold the perspective [15] that architectural specialization is not really necessary and uses a vanilla DINO as our backbone.

2.2 Panoramic Depth Estimation

With much larger view range than pinhole cameras, panoramic depth estimation serves as a promising technique for cost-efficient depth prediction method that may be used in autonomous driving and embodied AI systems. Compared with standard perspective imagery, panoramic images, which are most commonly represented in the equirectangular projection (ERP), introduce strong non-uniform distortion and a periodic discontinuity at the left–right image boundary. These properties encourages many unique design choices for panoramic depth estimation [2, 4, 17, 21–23, 29, 34, 38, 42, 48].

Specifically, Depth Any Camera (DAC) proposes a zero-shot metric depth estimation framework that extends a perspective-trained model to handle cameras with varying fields of view, such as fisheye and 360-degree cameras [10], but the robustness is still limited by data. To address the challenge of data scarcity, DA^2 [13] uses a data curation engine to generate high-quality panoramic depth data from perspective images. The recent DAP [16] further leverages a large-scale dataset created by combining public data, synthetic data, and real-world images to build a foundation model for panoramic metric depth estimation.

However, these models solely focus on monocular panoramic depth estimation that without known camera poses, reconstruction using these depth priors would be impossible, stressing the need for panoramic pose estimation.

2.3 Multi-view Panoramas

While multi-view 3D of pinhole cameras has been thoroughly discovered, researches into multi-view panoramas still have a long way to go.

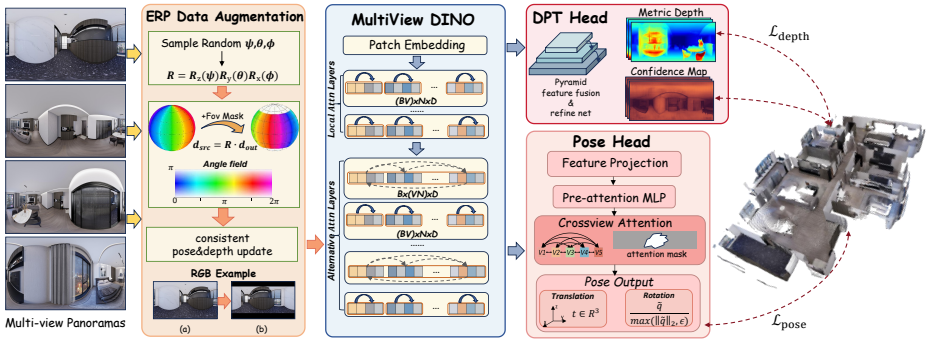


Fig. 3: The pipeline of CasaMaestro. From left to right we illustrate ERP data augmentation, model backbone, functional heads and training objective.

Many recent methods focus directly on panoramic 3DGS reconstruction [1,5, 27], achieving photorealistic reconstruction. However, they mainly operate on 2-5 views and are no built to be geometry-aware. In order to reconstruct longer sequences, others [32,35,45,46] uses dedicated pose estimation network for panoramas that either solve relative pose estimation between image pairs or directly process on scene-level. Nevertheless, these methods still suffer from the requirements of large view overlap, leaving the problem of house-level reconstruction efficiency yet to be solved.

3 Method

Fig. 3 presents the overall architecture of the proposed CasaMaestro, which employs a multi-view DINO backbone to extract dense features from an arbitrary number of input panoramas. Leveraging these shared representations, the dedicated DPT and pose heads simultaneously estimate depth maps and camera extrinsics to facilitate scene point cloud back-projection. We elaborate on the specific design of these components below.

3.1 Unified Backbone for Multi-view Representation Learning

To ensure an efficient model design, our backbone adopts the minimalist principle of utilizing a single plain ViT as in DepthAnything3 [15]. We construct this foundation upon vanilla DINOv2 [19] and facilitate multi-view reasoning via the input-adaptive cross-view self-attention mechanism, which is efficiently executed by reordering tokens during the forward pass. The backbone is configured to process exclusively visual inputs as the goal is to solve indoor panoramic reconstruction without known camera extrinsics.

Tokenization. In our formulation, each individual panoramic image serves as a distinct view. Assuming an arbitrary number of such views V for each sample,

we represent the input images as $\mathbf{I} \in \mathbb{R}^{B\times V\times 3\times H\times W}$. We then apply a standard patch embedding to each view to obtain the patch tokens $\mathbf{X}^{(0)} \in \mathbb{R}^{B\times V\times N\times D}$, where B stands for the batch size, N represents the number of patches per view, and D indicates the token dimension. We omit the class token for simplicity, although the formulation extends trivially if it is included. We use $\mathbf{X}^{(\ell)}$ to denote the token tensor produced after the ℓ-th transformer block.

Intra-view self-attention. During the first L_{local} layers, we perform self-attention independently within each individual view. To implement this operation concretely, we reshape the tokens as $\mathcal{R}(\mathbf{X}^{(\ell)}) \to \mathbb{R}^{(BV)\times N\times D}$ and subsequently apply a standard transformer block,

$$\mathbf{X}^{(\ell+1)} = \text{TransBlock}_{\text{local}}\left(\mathbf{X}^{(\ell)}\right), \qquad \ell < L_{\text{local}}. \tag{1}$$

By doing so, this initial stage effectively maintains the representation learning for each independent view and ensures that the early layers operate exactly like a standard transformer designed for single images.

Cross-view self-attention. Beginning at layer L_{local}, the backbone alternates between local (intra-view) attention within individual views and global (cross-view) attention across multiple views. We achieve this alternation *without* introducing any new attention layers. To implement this mechanism, we group tokens into distinct attention sequences. During the "local step", the model attends within each view as described in Eq. (1). Conversely, the "global step" attends across all views by flattening the view dimension directly into the token sequence. More specifically, for the global step, the tokens are reordered from $\mathbb{R}^{B\times V\times N\times D}$ to $\mathbb{R}^{B\times (VN)\times D}$. This resulting tensor is denoted as $\tilde{\mathbf{X}}^{(\ell)}$. Following this, we apply a standard transformer block over the sequence of length VN. We then reshape the resulting tensor back to its original dimensions,

$$\tilde{\mathbf{X}}^{(\ell+1)} = \text{TransBlock}_{\text{global}}\left(\tilde{\mathbf{X}}^{(\ell)}\right), \tag{2}$$

$$\mathbf{X}^{(\ell+1)} = \text{restore}\left(\tilde{\mathbf{X}}^{(\ell+1)}\right) \in \mathbb{R}^{B\times V\times N\times D}. \tag{3}$$

This backbone architecture effectively unifies local and global reasoning within a single transformer. The local attention mechanism maintains strong feature extraction for each individual view. Concurrently, the interleaved global steps allow tokens from distinct views to directly exchange information. This global interaction injects comprehensive context across the entire scene and significantly improves robustness even when viewpoints differ by large displacement.

We use standard DPT [25, 26] for per-view depth prediction, where multi-scale intermediate features are extracted for refinenet processing [15]. For camera pose head, we use the final output feature of backbone as input. The backbone is camera-agnostic by default and supports an arbitrary number of input views with a single set of weights.

3.2 Panoramic Camera Pose Decoder

Problem setup. Given the features extracted for each individual view by the upstream backbone, we decode the camera pose parameters for panoramic / wide-FoV settings. We achieve this by employing a transformer decoder head that operates across multiple views. Let $\mathbf{F} \in \mathbb{R}^{B\times V\times C}$ denote the input features, where C represents the feature dimension. The decoder output for each view v is defined as follows.

$$\hat{\mathbf{p}}_{b,v} = \left[\hat{\mathbf{t}}_{b,v},\ \hat{\mathbf{q}}_{b,v}\right] \in \mathbb{R}^7, \tag{4}$$

Here, $\hat{\mathbf{t}}_{b,v} \in \mathbb{R}^3$ represents the translation component, and $\hat{\mathbf{q}}_{b,v} \in \mathbb{R}^4$ denotes a unit quaternion rotation.

Motivation. A naive camera head can regress each view independently with an MLP, *i.e.*, $\hat{\mathbf{p}}_{b,v} = g(\mathbf{f}_{b,v})$ [15]. However, for panoramic / wide-FoV inputs containing multiple images, this independent regression lacks explicit information exchange across views. While the upstream backbone already incorporates global attention, its primary role is to extract versatile features suitable for dense depth prediction. Camera pose estimation is an inherently global geometric task that requires aligning fully condensed representations from multiple perspectives. Forcing the backbone to resolve these specific viewpoint ambiguities could compromise its dense feature extraction capabilities. Therefore, to explicitly enforce global consistency and decouple this geometric alignment from the main backbone, we introduce a panoramic camera pose head. This head applies attention across views during the decoding stage, as illustrated in the bottom right of Fig. 3. We provide validation of this design in Sec. 4.4

Feature projection. We first project $\mathbf{F}$ to the model dimension D with a linear layer followed by LayerNorm:

$$\mathbf{X}_0 = \mathrm{LN}(\mathbf{F}\mathbf{W}_p + \mathbf{b}_p) \in \mathbb{R}^{B\times V\times D}. \tag{5}$$

Per-view refinement (pre-attention). Before cross-view attention, we apply a per-view MLP with residual connection and LayerNorm to improve token conditioning:

$$\mathrm{MLP}(\mathbf{x}) = \mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2,\ \ \sigma(\cdot) = \mathrm{GELU}(\cdot), \tag{6}$$

$$\mathbf{X}_1 = \mathrm{LN}(\mathbf{X}_0 + \mathrm{MLP}(\mathbf{X}_0))\,. \tag{7}$$

The MLP hidden dimension is set to rD (with ratio r).

Cross-view self-attention. We then aggregate information across views through attention operated on the view dimension:

$$\mathbf{X}_2 = \mathrm{Attn}(\mathbf{X}_1;\ \mathbf{M}) \in \mathbb{R}^{B\times V\times D}, \tag{8}$$

where $\mathbf{M}$ is an optional attention mask to handle invalid regions in real-world captures (*e.g.*, camera poles or photographers needed to be masked out). A final LayerNorm produces the fused tokens:

$$\mathbf{Z} = \mathrm{LN}(\mathbf{X}_2). \tag{9}$$

Pose heads. Translation is regressed by a linear head:

$$\hat{\mathbf{t}}_{b,v} = \mathbf{Z}_{b,v}\mathbf{W}_t + \mathbf{b}_t \in \mathbb{R}^3. \tag{10}$$

Rotation is parameterized by a quaternion predicted by a linear head followed by explicit normalization:

$$\tilde{\mathbf{q}}_{b,v} = \mathbf{Z}_{b,v}\mathbf{W}_q + \mathbf{b}_q \in \mathbb{R}^4, \tag{11}$$

$$\hat{\mathbf{q}}_{b,v} = \frac{\tilde{\mathbf{q}}_{b,v}}{\max(\|\tilde{\mathbf{q}}_{b,v}\|_2,\ \epsilon)}. \tag{12}$$

This enforces geometrically valid rotations and improves numerical stability during training. For compatibility with the original DA3 behavior, if an external camera encoding $\mathbf{e}_{b,v}$ is provided, we optionally bypass the rotation regression and use its quaternion component $\hat{\mathbf{q}}_{b,v} = \mathrm{Norm}(\mathbf{e}_{b,v}[3{:}7])$.

While pinhole camera models may further estimate FOV values in $\mathbb{R}^2$, the standard panoramas come with fixed Field of View ($HFOV = 360°, VFOV = 180°$) and is hence free of prediction of such attributes, bringing a result of our pose decoder as in Eq. (4).

3.3 ERP Data Augmentation for Panoramas

Motivation. In the training set from current available datasets, viewpoints within the same scene often exhibit little to no relative rotation (*e.g.*, nearly identical yaw). This deficiency weakens the capacity of the model to develop reasoning capabilities regarding rotation, thereby hindering its generalization to diverse camera orientations. To address this limitation, we introduce an explicit rotation augmentation for equirectangular panoramic data equipped with paired RGB images, depth maps, and extrinsic parameters. This strategy produces additional relative rotations while strictly preserving the underlying 3D geometry.

Sampling random rotations. For each view in synthetic scenes, we sample a random rotation parameterized by yaw–pitch–roll:

$$\psi \sim \mathcal{U}(0,\ \psi_{\max}), \tag{13}$$

$$\theta \sim \mathrm{clip}\big(\mathcal{N}(0, \sigma_\theta^2),\ -\theta_{\max},\ \theta_{\max}\big), \tag{14}$$

$$\phi \sim \mathrm{clip}\big(\mathcal{N}(0, \sigma_\phi^2),\ -\phi_{\max},\ \phi_{\max}\big), \tag{15}$$

where ψ is yaw, θ is pitch, and ϕ is roll. We then construct a rotation matrix

$$\mathbf{R} = \mathbf{R}_{o_1}\mathbf{R}_{o_2}\mathbf{R}_{o_3}, \tag{16}$$

where (o_1, o_2, o_3) is a configurable multiplication order and $\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z$ are standard axis-angle rotations. In practice, (o_1, o_2, o_3) is defined as (z, y, x), and $\psi_{\max}, \theta_{\max}, \phi_{\max}$ are set to π, $\frac{\pi}{12}$, $\frac{\pi}{12}$ aligned with evaluation.

Equirectangular remapping. Given an equirectangular panorama of size $H \times W$, we associate each output pixel (u, v) with spherical angles

$$\theta = 2\pi \left(\frac{u}{W} - \frac{1}{2} \right), \qquad \varphi = \pi \left(\frac{1}{2} - \frac{v}{H} \right), \tag{17}$$

and convert to a 3D unit ray direction (using our panorama-ray convention):

$$\mathbf{d}_{\text{out}}(\theta, \varphi) = \begin{bmatrix} \cos\varphi \ \sin\theta \\ -\sin\varphi \\ \cos\varphi \ \cos\theta \end{bmatrix}. \tag{18}$$

To synthesize a rotated panorama, we rotate rays and sample from the source panorama. With our implementation convention,

$$\mathbf{d}_{\text{src}} = \mathbf{R} \, \mathbf{d}_{\text{out}}. \tag{19}$$

We then convert $\mathbf{d}_{\text{src}} = [x_s, y_s, z_s]^\top$ back to spherical angles

$$\theta_s = \text{atan2}(x_s, z_s), \qquad \varphi_s = \arcsin\big(\text{clip}(y_s, -1, 1)\big), \tag{20}$$

and finally obtain the sampling coordinates in the source image:

$$u_s = \left(\frac{\theta_s}{2\pi} + \frac{1}{2} \right) W, \qquad v_s = \left(\frac{1}{2} + \frac{\varphi_s}{\pi} \right) H. \tag{21}$$

We use bilinear interpolation and border wrap horizontally to respect the periodicity of panoramas.

Consistent pose update. Alongside warping the RGB and depth maps, we update the camera rotation in the extrinsic matrix to remain consistent with the augmented image. Let $\mathbf{T}_{\text{c2w}} \in \mathbb{R}^{4\times4}$ be the camera-to-world transform. We update its rotation block by right-multiplication:

$$\mathbf{R}_{\text{c2w}} \leftarrow \mathbf{R}_{\text{c2w}} \, \mathbf{R}. \tag{22}$$

The same remapping is applied to both RGB and depth to preserve pixel-wise alignment after augmentation. When a binary panorama mask is available, we apply it after the rotation warp (setting invalid pixels to zero), ensuring the masking operation remains consistent with the final augmented observations.

This augmentation injects controlled relative rotations into training scenes where native viewpoint rotations are limited, improving the model's robustness to orientation changes and encouraging cross-view reasoning under non-trivial camera rotations.

4 Experiment

4.1 Implementation Details

Our model is implemented using PyTorch framework and Adam optimizer. CasaMaestro is trained for 20 epoches on 4× NVIDIA A100 GPU with a per-GPU

Table 1: Quantitative pose metrics comparison on Realsee-Real. The $*$ denotes finetune or re-implementation on Realsee training set, and $^\circ$ indicates panoramic methods.

Method	AUC@10 ↑	AUC@20 ↑	AUC@30 ↑	Rot. Mean ↓	Trans. Mean ↓	Pose Mean ↓
PanoPose*° [32]	0.042	0.115	0.166	62.20	58.27	90.12
SPR*° [46]	0.086	0.145	0.191	59.43	56.12	86.23
VGGT [37]	0.189	0.313	0.388	37.74	32.31	47.65
VGGT-Finetune*	0.276	0.379	0.411	30.12	10.23	38.42
PI3 [41]	0.222	0.365	0.450	26.69	25.48	36.61
PI3-Finetune*	0.301	0.397	0.510	21.36	7.442	28.97
DepthAnything3 [15]	0.083	0.173	0.235	54.57	49.79	69.51
StreamVGGT [47]	0.033	0.039	0.062	52.46	5.351	70.58
InfiniteVGGT [44]	0.271	0.412	0.487	29.82	4.488	30.01
CasaMaestro°(Ours)	**0.727**	**0.859**	**0.903**	**2.608**	**2.132**	**3.225**

Table 2: Quantitative pose metrics comparison on Realsee-Syn.

Method	AUC@10 ↑	AUC@20 ↑	AUC@30 ↑	Rot. Mean ↓	Trans. Mean ↓	Pose Mean ↓
PanoPose*° [32]	0.082	0.135	0.179	57.44	57.39	88.04
SPR*° [46]	0.093	0.162	0.201	51.24	50.16	67.42
VGGT [37]	0.118	0.226	0.292	48.93	42.99	61.49
VGGT-Finetune*	0.244	0.386	0.423	29.20	9.672	30.78
PI3 [41]	0.237	0.347	0.410	38.82	21.77	28.65
PI3-Finetune*	0.310	0.412	0.507	23.43	5.320	25.31
DepthAnything3 [15]	0.030	0.073	0.110	71.24	59.67	87.20
StreamVGGT [47]	0.007	0.007	0.010	57.84	11.56	102.0
InfiniteVGGT [44]	0.068	0.196	0.288	61.37	6.991	61.83
CasaMaestro°(Ours)	**0.792**	**0.892**	**0.927**	**1.553**	**1.682**	**2.281**

batchsize of 1 and a learning rate of 5e-5. Input view number ranges from 20 to 50, and the DINO backbone is initialized from DA3 [15]. Evaluation is conducted on a single NVIDIA A100 GPU. Training and evaluation resolution are set to 448 for single side with fixed aspect ratio as in DA3 [15]. For pinhole methods, the data is split into pinhole images through ERP projection with $\frac{\pi}{2}$ yaw step (results are stable with different yaw steps as shown in Fig. 2). All comparing methods are configured following their default settings. CaseMaestro is trained on **Realsee3D** [14] dataset, and we also evaluate the robustness of CaseMaestro by conducting zero-shot evaluation on other multi-view indoor panoramic datasets including **PanoSUNCG** [33], **The Habitat-Matterport 3D Research Dataset (HM3D)** [24] and **Replica** [30].

4.2 Quantitative Experiments

We conduct quantitative experiments on the sub-tasks of 3D point cloud reconstruction including camera pose estimation and depth estimation.

Camera Pose Estimation. We report the results of AUC and pose error of the Realsee dataset (test set) in Tab. 1 and Tab. 2, where pose error is defined as

Table 3: Quantitative depth estimation comparison on Realsee dataset.

Method	Data Split: Real-World			Data Split: Synthetic			Data Split: Overall		
	AbsRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	RMSE ↓	δ_1 ↑
PanoPose*° [32]	0.143	0.504	0.801	0.156	0.319	0.773	0.155	0.336	0.775
DAP° [16]	0.144	0.530	0.809	0.188	0.406	0.723	0.184	0.417	0.731
VGGT [37]	0.159	0.572	0.772	0.163	0.332	0.751	0.162	0.354	0.753
InfiniteVGGT [44]	0.159	0.381	0.756	0.189	0.361	0.697	0.186	0.363	0.702
Pi3 [41]	0.423	0.792	0.031	0.140	0.308	0.802	0.166	0.352	0.732
DepthAnything3 [15]	0.203	0.617	0.705	0.153	0.325	0.745	0.157	0.351	0.741
CasaMaestro°(Ours)	**0.078**	**0.482**	**0.940**	**0.079**	**0.183**	**0.975**	**0.078**	**0.205**	**0.972**

the maximum error of rotation error and translation error. As shown, previous panoramic methods PanoPose and SPR rely highly on view overlap fail largely even when re-implemented on Realsee dataset, because these methods are designed for video streams with dense input frames. Foundation models show relatively better quality due to large pre-training but are not stable as the metrics of both DepthAnything3 and InfiniteVGGT largely drop when it comes to synthetic scenes. VGGT and Pi3 finetuned on Realsee training set is stable but fails to acquire significant improvements. InfiniteVGGT and StreamVGGT both have low translation error, but still faces a large rotation error when applied to inputs with little overlap instead of video streams. CasaMaestro stands the only one with the ability to correctly restore house structures, outperforming these methods with **84%** and **119%** improvements of the AUC30 metric than the best previous results in real-world and synthetic scenes with balanced rotation and translation error.

Depth Estimation. We first compare the depth estimation quality on Realsee dataset as shown in Tab. 3. Although PanoPose shows good quality, it is re-implemented on Realsee training set and common VFMs all provide zero-shot results with similar quality. CasaMaestro significantly outperforms these models on Realsee dataset, yet while depth estimation could be data dependent and comparing with pinhole VFMs may not be fair, we further showcase the zero-shot depth estimation ability of CasaMaestro in three unseen datasets comparing only with panoramic depth estimation methods. As shown in Tab. 4, even compared to dedicated single-view panoramic depth estimation methods or models that predict relative depth, the metric depth predicted by CasaMaestro is still superior, specifically an average **21.98%** decrease in AbsRel compared to the best previous results, while no modules are intentionally designed for this task.

The above pose and depth estimation results demonstrate the robustness of CasaMaestro's panoramic 3D reconstruction ability across different scenes.

4.3 Qualitative Comparison

We provide qualitative visual comparison in Fig. 4 and Fig. 5. As shown, in circumstances of sparse viewpoints, all existing methods face severe errors and

Table 4: Quantitative zero-shot depth estimation comparison.

Method	Dataset								
	Replica			PanoSUNCG			HM3D		
	AbsRel ↓	RMSElog ↓	δ_1 ↑	AbsRel ↓	RMSElog ↓	δ_1 ↑	AbsRel ↓	RMSElog ↓	δ_1 ↑
Single-View Methods									
PanoFormer [29]	0.076	0.129	0.945	0.054	0.122	0.978	0.151	0.184	0.742
DA^2 [13]	0.070	0.091	0.966	0.060	0.181	0.976	0.164	0.195	0.751
DAP [16]	0.124	0.161	0.878	0.130	0.169	0.841	0.143	0.216	0.749
Multi-View Methods									
PanoPose* [32]	0.116	0.134	0.937	0.094	0.142	0.955	0.215	0.226	0.708
CasaMaestro (Ours)	**0.058**	**0.086**	**0.970**	**0.042**	**0.116**	**0.988**	**0.105**	**0.130**	**0.921**

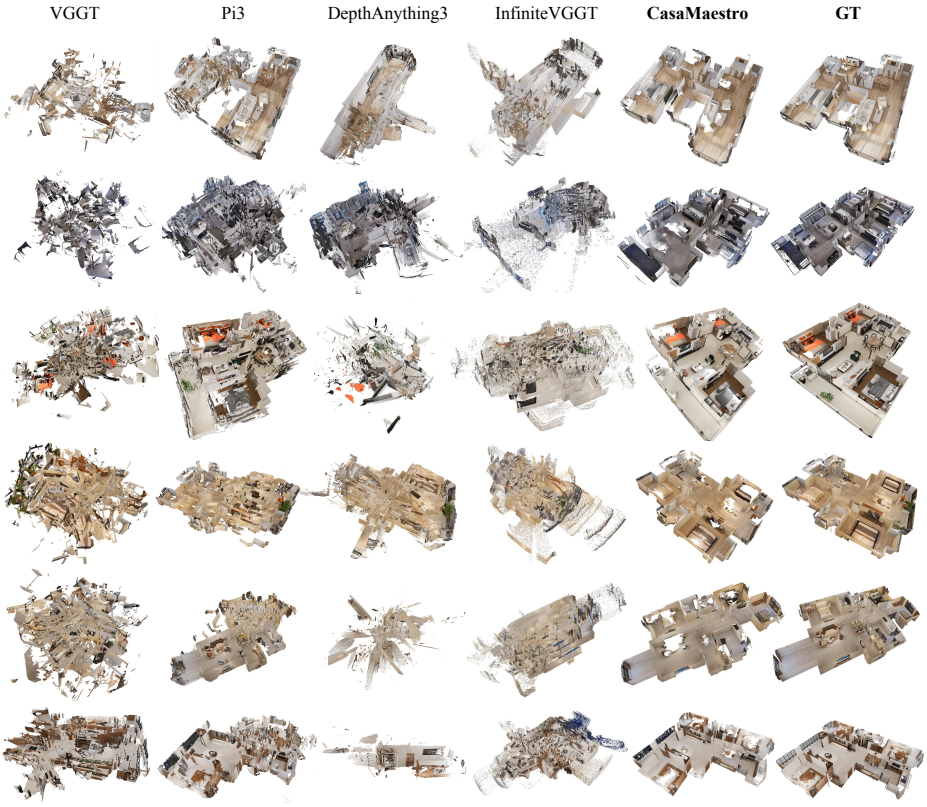


Fig. 4: Quantitative comparison on Realsee-Syn. Upper 50% points are filtered to remove the rooftop for visibility.

fail to generate reasonable reconstructions. This result is in accord with metrics reported in Tab. 1 and Tab. 2. Specifically, although InfiniteVGGT shows the least translation error among existing methods, the rotation error is high, which still leads to much pose error especially in synthetic scenes and the resulting

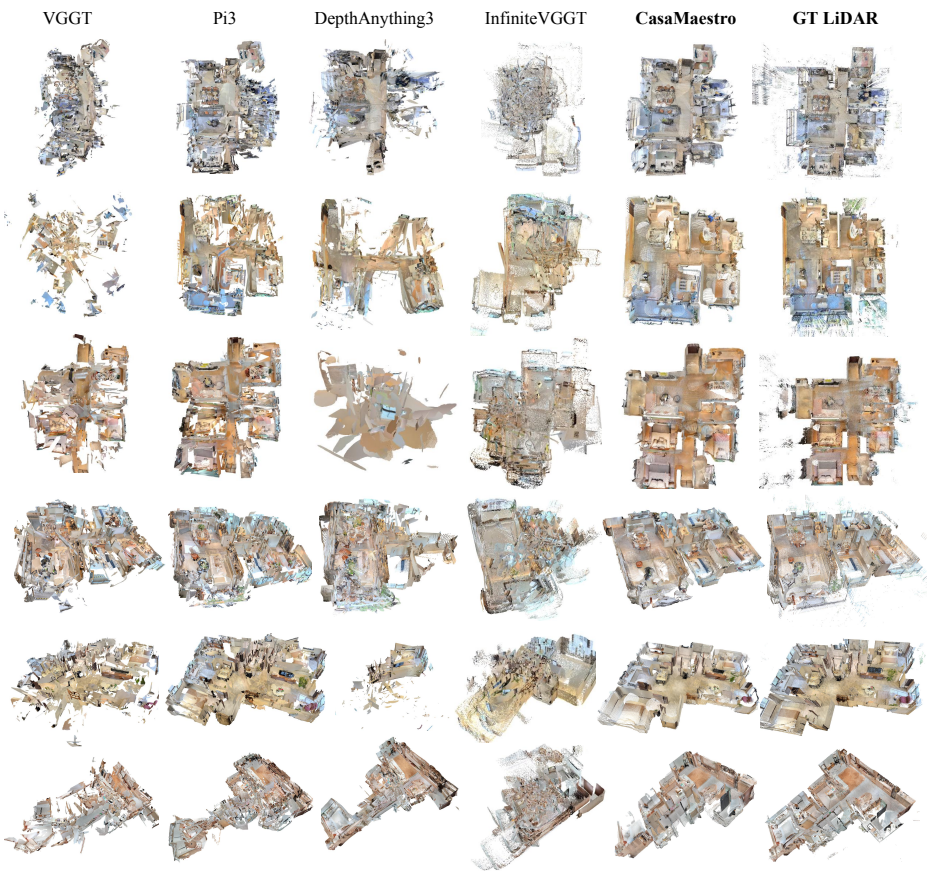


Fig. 5: Quantitative comparison on Realsee-Real.

reconstruction becomes messy. Pi3 has more balanced rotation error and translation error, providing more reasonably reconstructed scenes, yet still creates much noise. Meanwhile, VGGT and DepthAnything face more cases of complete corruption in certain scenes than others, standing for the high pose errors. In the figures, CasaMaestro shows first-tier ability of accurately reconstructing houses with multiple rooms.

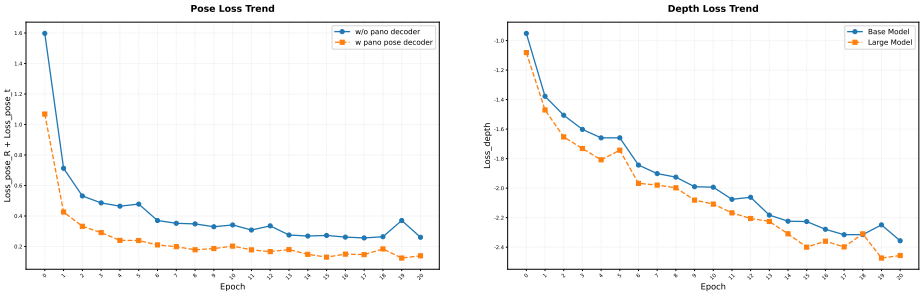
4.4 Ablation Study

We conduct ablation to validate the rationality of our design. The ablated components are model backbone, our panoramic camera pose decoder and ERP data augmentation strategy.

We provide comparison of final metrics in Tab. 5. As reported, simply using a larger backbone does not show significant difference (① *v.s.* ⑤), which is in accordance with the trend of train-time depth loss shown in Fig. 6b, indicating

Table 5: Ablation study with choices of key model components.

#	Components					Pose Metrics						Depth Metrics		
	$DINO_{base}$	$DINO_{large}$	$PoseDec_{Linear}$	$PoseDec_{Ours}$	$ERP_{Aug.}$	AUC@10 ↑	AUC@20 ↑	AUC@30 ↑	Rot. Mean ↓	Trans. Mean ↓	Pose Mean ↓	AbsRel ↓	RMSElog ↓	δ_1 ↑
①	✓		✓			0.302	0.454	0.568	24.39	17.85	28.11	0.092	0.110	0.920
②	✓			✓		0.629	0.808	0.831	3.192	2.834	4.021	0.077	0.104	0.939
③	✓		✓		✓	0.512	0.651	0.772	2.175	2.375	2.989	0.111	0.125	0.970
④	✓			✓	✓	0.688	0.839	0.891	2.547	2.404	3.380	0.075	0.097	0.956
⑤		✓	✓			0.297	0.456	0.598	24.41	16.32	25.43	0.077	0.097	0.939
⑥		✓		✓		0.707	0.829	0.871	2.102	2.175	2.837	0.076	0.105	0.973
⑦		✓	✓		✓	0.604	0.719	0.815	2.132	2.265	2.774	0.080	0.114	0.938
⑧		✓		✓	✓	**0.792**	**0.892**	**0.927**	**1.553**	**1.682**	**2.281**	0.078	0.103	0.975



(a) Ablation on pose decoder. **(b)** Ablation on model size.

Fig. 6: Ablations with comparison of loss trends. Loss in epoch 0 means the average loss of the first epoch.

that this problem can not be solved simply during feature extraction stage. Using our panoramic camera pose decoder provides more improvements comparing with solely applying ERP data augmentation (②,⑥ *v.s.* ③,⑦). Meanwhile, as shown in Fig. 6a, during training, using the panoramic camera pose decoder will have much faster convergence than common linear decoder that takes only 3 epoches to reach the quality of 10+ epoches without it. This shows that the extra attention applied to our proposed pose decoder indeed help with pose prediction instead of providing only metric improvements of the final results.

The above experiments demonstrate CasaMaestro's first-tier performance and the design rationality. For more information including performance report and other experiments, please refer to the *Supplementary Material*.

5 Conclusion

In this work, we present CasaMaestro, the first feedforward model that achieves extrinsic-free multi-view panoramic 3D reconstruction with sparse house-scale capture. CasaMaestro significantly outperforms existing state-of-the-art models with an excellent ability to restore the house structure, while also showing a strong zero-shot ability of metric depth on unseen datasets. The results prove that feedforward any-view 3D reconstruction needs not to be restricted to pinhole cameras, and the rise of more Vision Foundation Models built upon panoramas and fisheye cameras could provide a new revolution in various 3D tasks.

References

1. Bai, J., Huang, L., Guo, J., Gong, W., Li, Y., Guo, Y.: 360-gs: Layout-guided panoramic gaussian splatting for indoor roaming. In: 2025 International Conference on 3D Vision (3DV). pp. 1042–1053. IEEE (2025)
2. Bai, J., Qin, H., Lai, S., Guo, J., Guo, Y.: Glpanodepth: Global-to-local panoramic depth estimation. IEEE Transactions on Image Processing **33**, 2936–2949 (2024)
3. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1050–1060 (2025)
4. Cao, Z., Zhu, J., Zhang, W., Ai, H., Bai, H., Zhao, H., Wang, L.: Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 982–992 (2025)
5. Chen, Z., Wu, C., Shen, Z., Zhao, C., Ye, W., Feng, H., Ding, E., Zhang, S.H.: Splatter-360: Generalizable 360 gaussian splatting for wide-baseline panoramic images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21590–21599 (2025)
6. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences (2025), https://arxiv.org/abs/2507.16443
7. Dong, S., Wang, S., Liu, S., Cai, L., Fan, Q., Kannala, J., Yang, Y.: Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. arXiv preprint arXiv:2412.08376 (2024)
8. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In: 2025 International Conference on 3D Vision (3DV). pp. 1–10. IEEE (2025)
9. Elflein, S., Zhou, Q., Leal-Taixé, L.: Light3r-sfm: Towards feed-forward structure-from-motion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16774–16784 (2025)
10. Guo, Y., Garg, S., Miangoleh, S.M.H., Huang, X., Ren, L.: Depth any camera: Zero-shot metric depth estimation from any camera. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26996–27006 (2025)
11. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P.: MapAnything: Universal feed-forward metric 3D reconstruction (2025), arXiv preprint arXiv:2509.13414
12. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024)
13. Li, H., Zheng, W., He, J., Liu, Y., Lin, X., Yang, X., Chen, Y.C., Guo, C.: Da ^{2}: Depth anything in any direction. arXiv preprint arXiv:2509.26618 (2025)
14. Li, L., Wu, Y., Li, X., Wang, L., Rao, T., Zhou, J., Pan, C., Hui, X.: Realsee3d: A large-scale multi-view rgb-d dataset of indoor scenes (version 1.0) (2025). https://doi.org/10.5281/zenodo.17826243, https://doi.org/10.5281/zenodo.17826243
15. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)

16. Lin, X., Song, M., Zhang, D., Lu, W., Li, H., Du, B., Yang, M.H., Nguyen, T., Qi, L.: Depth any panoramas: A foundation model for panoramic depth estimation. arXiv preprint arXiv:2512.16913 (2025)
17. Liu, J., Xu, Y., Li, S., Li, J.: Estimating depth of monocular panoramic image with teacher-student model fusing equirectangular and spherical representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1262–1271 (2024)
18. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16695–16705 (2025)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
20. Pataki, Z., Sarlin, P.E., Schönberger, J.L., Pollefeys, M.: Mp-sfm: Monocular surface priors for robust structure-from-motion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21891–21901 (2025)
21. Peng, C.H., Zhang, J.: High-resolution depth estimation for 360-degree panoramas through perspective and panoramic depth images registration. arXiv preprint arXiv:2210.10414 (2022)
22. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: UniK3D: Universal camera monocular 3d estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
23. Pintore, G., Agus, M., Almansa, E., Schneider, J., Gobbetti, E.: Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11536–11545 (2021)
24. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
25. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ArXiv preprint (2021)
26. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
27. Ren, J., Xiang, M., Zhu, J., Dai, Y.: Panosplatt3r: Leveraging perspective pretraining for generalized unposed wide-baseline panorama reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28959–28969 (2025)
28. Shen, Y., Zhang, Z., Qu, Y., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
29. Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: Panoformer: panorama transformer for indoor 360° depth estimation. In: European Conference on Computer Vision. pp. 195–211. Springer (2022)
30. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)

31. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. arXiv preprint arXiv:2412.06974 (2024)
32. Tu, D., Cui, H., Zheng, X., Shen, S.: Panopose: Self-supervised relative pose estimation for panoramic images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20009–20018 (2024)
33. Wang, F.E., Hu, H.N., Cheng, H.T., Lin, J.T., Yang, S.T., Shih, M.L., Chu, H.K., Sun, M.: Self-supervised learning of depth and camera motion from 360 videos. In: Asian Conference on Computer Vision. pp. 53–68. Springer (2018)
34. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 462–471 (2020)
35. Wang, F.E., Yeh, Y.H., Tsai, Y.H., Chiu, W.C., Sun, M.: Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5448–5460 (2022)
36. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024)
37. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
38. Wang, N.H.A., Liu, Y.L.: Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. Advances in Neural Information Processing Systems **37**, 127739–127764 (2024)
39. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
40. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
41. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Scalable permutation-equivariant visual geometry learning (2025), https://arxiv.org/abs/2507.13347
42. Yan, Z., Li, X., Wang, K., Zhang, Z., Li, J., Yang, J.: Multi-modal masked pre-training for monocular panoramic depth completion. In: European Conference on Computer Vision. pp. 378–395. Springer (2022)
43. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2025)
44. Yuan, S., Yang, Y., Yang, X., Zhang, X., Zhao, Z., Zhang, L., Zhang, Z.: Infinitevggt: Visual geometry grounded transformer for endless streams (2026)
45. Yun, I., Lee, H.J., Rhee, C.E.: Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3224–3233 (2022)
46. Zheng, J., Liu, R., Chen, Y., Chen, Z., Yang, K., Zhang, J., Stiefelhagen, R.: Scene-agnostic pose regression for visual localization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27092–27102 (2025)
47. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)

48. Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 448–465 (2018)