HunyuanWorld-Mirror: Technical Report

Tencent Hunyuan*

https://3d-models.hunyuan.tencent.com/world/
https://huggingface.co/tencent/HunyuanWorld-Mirror
https://github.com/Tencent-Hunyuan/HunyuanWorld-Mirror

Abstract

While **HunyuanWorld 1.0** generates immersive and playable 3D worlds from texts or single-view images, it lacks the capability to process videos or multi-view images. **HunyuanWorld 1.1** bridges this gap with *WorldMirror*, an all-in-one, feed-forward model for versatile 3D geometric prediction tasks, which unlocks video-to-3D and multi-view-to-3D world creation. Unlike existing methods constrained to image-only inputs or customized for a specific task, our framework flexibly integrates diverse geometric priors, including camera poses, intrinsics, and depth maps, while simultaneously generating multiple 3D representations: dense point clouds, multi-view depth maps, camera parameters, surface normals, and 3D Gaussians. This elegant and unified architecture leverages available prior information to resolve structural ambiguities and delivers geometrically consistent 3D outputs in a single forward pass. *WorldMirror* achieves state-of-the-art performance across diverse benchmarks from camera, point map, depth, and surface normal estimation to novel view synthesis, while maintaining the efficiency of feed-forward inference.

1 Introduction

Visual geometry learning is a fundamental problem in computer vision, with applications spanning augmented reality, robotics, and autonomous navigation. Traditional Structure-from-Motion (SfM) [41] and Multi-View Stereo (MVS) algorithms rely on iterative optimization, making them computationally expensive. The field has recently shifted toward feed-forward neural networks that directly reconstruct geometry from visual inputs. These end-to-end models, exemplified by DUSt3R [54] and its successors, have demonstrated remarkable capabilities in processing image pairs, videos, and multi-view images.

∗ Team contributors are listed in the end of report.

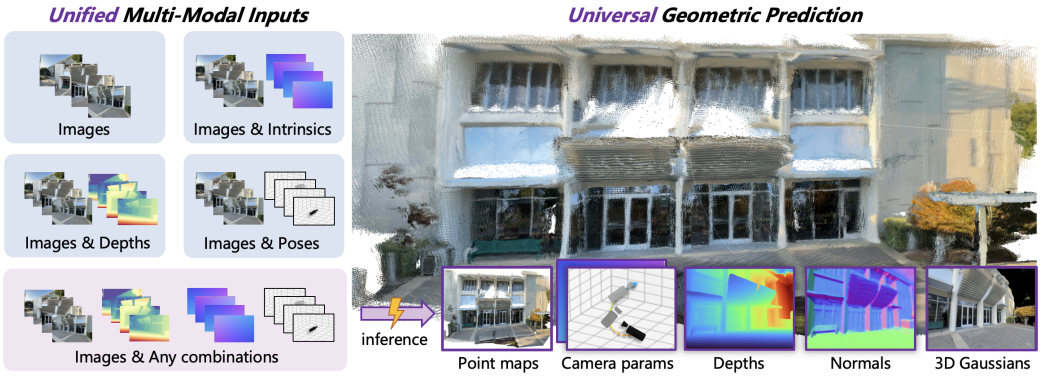

Figure 1: *HunyuanWorld-Mirror* is a large feed-forward 3D reconstruction model that takes raw images along with optional priors (depth, calibrated intrinsics, camera pose) as input and produces high-quality geometric attributes in seconds, including point clouds, 3DGS, cameras, depth, and normal maps.

Despite significant progress, existing methods still face two critical limitations regarding their input and output spaces. On the input front, these approaches exclusively process raw images, failing to leverage additional modalities that are useful and often accessible in real-world applications, such as calibrated camera intrinsics, camera poses, and depth measurements derived from LIDAR or RGB-D sensors. Without incorporating these prior cues, current methods encounter unnecessary challenges in scenarios that could otherwise be readily addressed: calibrated intrinsics resolve scale ambiguities, camera poses ensure multi-view consistency, and depth measurements ground predictions in areas where image-based cues alone are insufficient, such as textureless or reflective regions.

Second, existing methods are typically limited to addressing single or limited tasks in output space. These approaches are often highly specialized, *e.g.*, focusing on depth estimation [61], point map regression [54], camera pose prediction [51], or point tracking [24], and rarely integrate multiple tasks within a unified framework. Recently, VGGT [50] has explored unifying these tasks, but some fundamental geometry tasks like surface normal estimation and novel view synthesis remain excluded. These two limitations prompt a critical question: can we reconcile both challenges by effectively leveraging diverse prior knowledge within a universal 3D reconstruction architecture?

To address these challenges, we introduce *WorldMirror* [32], a framework designed to perform universal 3D reconstruction tasks while leveraging any available geometric priors. At the core of *WorldMirror* is a novel **Multi-Modal Prior Prompting** mechanism that embeds diverse prior modalities, including calibrated intrinsics, camera pose, and depth, into the feed-forward model. Given any subset of the available priors, we utilize several lightweight encoding layers to convert each modality into structured tokens. Rather than treating all prior modalities uniformly, we implement specialized embedding strategies for each modality type. Camera poses and calibrated intrinsics are encoded into a single token due to their compact nature. Depth maps, rich in spatial information, are converted to dense tokens. These tokens maintain spatial alignment with visual tokens and are integrated through direct addition. Furthermore, to reduce the training-inference gap, we propose a dynamic prior injection scheme by randomly sampling distinct prior combinations during training, enabling the model to adapt to arbitrary subsets (including none) of available priors during inference.

Besides, *WorldMirror* features a **Universal Geometric Prediction** architecture capable of handling the full spectrum of 3D reconstruction tasks from camera and depth estimation to point map regression, surface normal estimation, and novel view synthesis. *WorldMirror* builds upon a fully transformer-based architecture for regressing camera parameters and uses unified decoder heads for all other dense prediction tasks. Incorporating these tasks together broadens the model's capabilities toward a versatile 3D reconstruction framework. However, training such a multi-task 3D reconstruction foundation model poses significant challenges, as geometric quantities are inherently coupled and require carefully designed training strategies. We thus propose a systematic curriculum learning

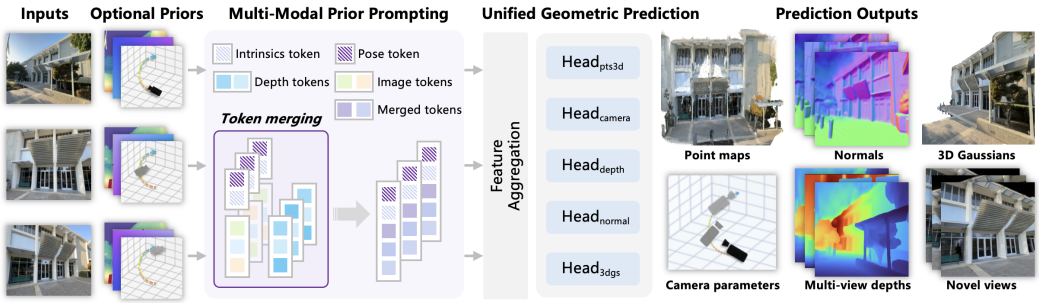

Figure 2: **Overview of *WorldMirror*.** Given multi-view images with optional priors (depths, calibrated intrinsics, camera poses) as input, our framework encodes each prior modality into tokens and integrates them with image tokens. The composite tokens are subsequently processed by a visual transformer backbone to effectively aggregate multi-view features. The consolidated representations are then passed to multi-task heads to generate comprehensive geometric outputs, including point maps, camera parameters, multi-view depth maps, surface normals, and 3D Gaussians.

strategy to optimize training efficiency and enhance performance by progressing from simple to complex across three dimensions: task sequencing, data scheduling, and progressive resolution.

Extensive experiments demonstrate that *WorldMirror* achieves state-of-the-art performance across diverse benchmarks and tasks. It surpasses recent 3D reconstruction methods, such as VGGT [50] and π^3 [56] in point map and camera estimation, while outperforming StableNormal [64] and GeoWizard [15] in surface normal prediction and significantly exceeding recent method AnySplat [22] in novel view synthesis.

Our contributions can be summarized as:

- We propose a universal 3D world reconstruction model capable of taking multi-modal priors as guidance, including per-view calibrated intrinsics, camera pose, and depth maps.
- Our model serves as a foundational 3D reconstruction framework, which supports universal geometric predictions from point map, camera, depth, and surface normal estimation to novel view synthesis.
- Extensive experiments show that our method outperforms existing methods across diverse tasks qualitatively and quantitatively.

2 Technical Details

Given N multi-view images $\{\boldsymbol{I}_i\}_{i=1}^N$, our work aims to utilize any available priors for unified geometric predictions. To this end, we introduce *multi-modal prior prompting* (Sec. 2.1) to embed priors including calibrated intrinsics, camera poses, and depth maps seamlessly into dense visual tokens as guidance for our model. To unify various geometric predictions, we present *universal geometric prediction* (Sec. 2.2) to predict various geometric attributes, including point maps, multi-view depths, camera parameters, surface normals, and 3D Gaussians, within our unified framework. To reduce the training-inference gap and achieve the optimal overall performance, we introduce a dynamic prior injection scheme with well-designed curriculum learning strategies (Sec. 3.1).

2.1 Multi-Modal Prior Prompting

As demonstrated in previous works [36, 19], auxiliary information like calibrated intrinsics, depths, and camera poses substantially enhances visual geometric learning. This motivates us to develop a model that flexibly leverages available priors when present, while maintaining robust reconstruction quality when priors are unavailable. In the following, we discuss how to effectively embed diverse modality information as input to our model, and then describe the training strategy that enables the model to flexibly infer with any priors.

Camera Pose. Given the camera poses $\{[\boldsymbol{R}_i|\boldsymbol{t}_i]\}_{i=1}^N$ of input images, where $\boldsymbol{R}_i \in \mathbb{R}^{3\times3}, \boldsymbol{t}_i \in \mathbb{R}^3$, we first normalize the scene scale to a standard unit cube, and the new translation vector $\boldsymbol{t}^{norm}$ is formulated as: $\boldsymbol{t}_i^{norm} = (\boldsymbol{t}_i - \boldsymbol{c})/\alpha$, where $\boldsymbol{c}$ is the camera center and α is the maximum distance of each camera to $\boldsymbol{c}$. This normalization ensures consistent numerical ranges regardless of the scene scale. Then, to integrate camera information, we encode each camera pose $[\boldsymbol{R}_i|\boldsymbol{t}_i^{norm}]$ into a single token due to their compact representation. Specifically, we convert each rotation matrix $\boldsymbol{R}_i \in \mathbb{R}^{3\times3}$ to a quaternion $\boldsymbol{q}_i \in \mathbb{R}^4$ and combine it with the normalized translation vector $\boldsymbol{t}_i^{norm} \in \mathbb{R}^3$ to form a 7-dimensional vector. This vector is then projected to $\boldsymbol{T}_i^{cam} \in \mathbb{R}^{1\times D}$ using a two-layer MLP, where D matches the dimension of image tokens, enabling seamless token concatenation.

Calibrated Intrinsics. Embedding calibrated camera intrinsics is comparatively straightforward. Given the intrinsic matrix $\boldsymbol{K}_i \in \mathbb{R}^{3\times3}$ of each image, we extract the focal lengths and principal points (f_x, f_y, c_x, c_y) and normalize them by dividing the image width W and height H, respectively. This normalization ensures training stability across images with varying resolutions. Similar to camera pose, we project the normalized intrinsic to $\boldsymbol{T}_i^{intr} \in \mathbb{R}^{1\times D}$ using a two-layer MLP, enabling seamless concatenation with visual tokens.

Depth Map. Unlike camera poses and intrinsics that are compact representations, depth maps are dense spatial signals requiring different embedding strategies. Given a depth map $\boldsymbol{D}_i \in \mathbb{R}^{H\times W}$, we first normalize its values to the range $[0, 1]$ to ensure numerical stability. Then, we employ a convolutional layer with kernel size matching the patch size used for visual tokens to create depth tokens $\boldsymbol{T}_i^{depth} \in \mathbb{R}^{(H_p\times W_p)\times D}$, where H_p, W_p are the token height and width, respectively. These depth tokens are spatially aligned with the visual tokens and are directly added to them. This additive integration preserves the spatial structure of the scene while enriching visual tokens with geometric information, fusing appearance and geometry in a unified representation.

Versatile Prior Prompting. To enable versatile prior-prompted 3D reconstruction, we concatenate intrinsics tokens and camera pose tokens with image tokens $\boldsymbol{T}_i^{img} \in \mathbb{R}^{(H_p\times W_p)\times D}$, while directly adding depth tokens, resulting in a prompted token set $\boldsymbol{T}_i^{prompt}$ as:

$$\boldsymbol{T}_i^{prompt} = [\boldsymbol{T}_i^{cam},\ \boldsymbol{T}_i^{intr}, \boldsymbol{T}_i^{img} + \boldsymbol{T}_i^{depth}], \quad \boldsymbol{T}_i^{prompt} \in \mathbb{R}^{(1+1+H_p\times W_p)\times D} \tag{1}$$

Considering that during inference, we may not have access to all modality information, we thus propose a dynamic prior injection scheme during training, which allows the model to adapt to arbitrary combinations of priors, as stated in Sec. 3.1.

2.2 Universal Geometric Prediction

Recent approaches, such as VGGT, have unified various geometry prediction tasks, but lack support for some common applications like novel view synthesis and surface normal estimation. In this work, we propose a more comprehensive framework enabling universal geometric prediction, including point maps, camera parameters, depth maps, surface normals, and 3D Gaussians.

Point Map, Camera, and Depth Estimation. Following the design of VGGT, given the output tokens $\boldsymbol{T}_i^{out} \in \mathbb{R}^{L\times D}$ of visual transformer backbone, we utilize DPT heads $\texttt{DPT}(\cdot)$ [37] to regress dense outputs, including 3D point map $\hat{\boldsymbol{P}}_i$ and multiview depth $\hat{\boldsymbol{D}}_i$, and use transformer layers to predict camera parameters $\hat{\boldsymbol{E}}_i$ from camera tokens:

$$\hat{\boldsymbol{P}}_i = \texttt{DPT}_p(\hat{\boldsymbol{T}}_i^{img}), \quad \hat{\boldsymbol{D}}_i = \texttt{DPT}_d(\hat{\boldsymbol{T}}_i^{img}), \quad \hat{\boldsymbol{E}}_i = \texttt{Transformer}(\hat{\boldsymbol{T}}_i^{cam}) \tag{2}$$

Surface Normal Estimation. For surface normal estimation, we employ the same DPT architecture as other dense prediction tasks, followed by L2 normalization to ensure unit vector outputs:

$$\hat{\boldsymbol{N}}_i = \texttt{DPT}_n(\hat{\boldsymbol{T}}_i^{img}) \ / \ ||\texttt{DPT}_n(\hat{\boldsymbol{T}}_i^{img})||_2. \tag{3}$$

To address the scarcity of ground-truth normal annotations, we introduce a hybrid supervision approach. We leverage both annotated datasets and pseudo normals derived from ground-truth depth maps via plane fitting for datasets lacking normal labels, which enables effective usage of diverse data for generalization while ensuring consistent normal estimation.

Novel View Synthesis. To enable novel view synthesis, we predict 3D Gaussian Splatting (3DGS). Specifically, we use a DPT head $\texttt{DPT}_g(\cdot)$ to regress pixel-wise Gaussian depth maps $\hat{\boldsymbol{D}}_g$ and Gaussian

feature maps F_g. These depth predictions are back-projected using the ground-truth camera poses $[\boldsymbol{R}|\boldsymbol{t}]$ and intrinsic matrix $\boldsymbol{K}$ to obtain the Gaussian centers $\boldsymbol{\mu}_g$. To infer the remaining Gaussian attributes $\hat{\boldsymbol{G}}$, including opacity σ_g, orientation r_g, scale s_g, residual spherical-harmonic color coefficients $\Delta\mathbf{c}_g$, and a fusion weight w_g, we combine F_g with appearance features derived from a convolution network $\texttt{Conv}(\cdot)$. The overall process can be formulated as:

$$\hat{\boldsymbol{G}} = \texttt{Conv}(F_g, \boldsymbol{I}), \qquad \hat{\boldsymbol{D}}_g, F_g = \texttt{DPT}_g(\hat{\boldsymbol{T}}^{img}) \tag{4}$$

To reduce Gaussian redundancy caused by overlapping regions across multiple views, we cluster and prune per-pixel Gaussians through voxelization, similar to AnySplat [22]. To enable novel view synthesis, the input images are split into context and target sets during training. The 3D Gaussians are built only from context views but rendered to and supervised by both target and original context viewpoints via a differentiable rasterizer [65]. This dual supervision enables the model to synthesize novel views while preserving consistency with input observations.

Training Losses. Our model is trained end-to-end by minimizing a composite loss function, $\mathcal{L}$, which integrates supervision for all prediction tasks:

$$\mathcal{L} = \lambda_{\text{points}}\mathcal{L}_{\text{points}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{3dgs}}\mathcal{L}_{\text{3dgs}}. \tag{5}$$

We follow VGGT to implement $\mathcal{L}_{\text{cam}}$, $\mathcal{L}_{\text{points}}$, and $\mathcal{L}_{\text{depth}}$. Specifically, we use a gradient-based term to supervise the predicted point $\hat{P}_i$:

$$\mathcal{L}_{\text{point}} = \sum_{i=1}^{N} \|\Sigma_i^P \odot (\hat{\boldsymbol{P}}_i - \boldsymbol{P}_i)\| + \|\Sigma_i^P \odot (\nabla\hat{\boldsymbol{P}}_i - \nabla\boldsymbol{P}_i)\| - \alpha \log \Sigma_i^P, \tag{6}$$

where $\odot$ is the channel-broadcast element-wise product and Σ_i^P refers to the point uncertainty. The depth loss $\mathcal{L}_{\text{depth}}$ is analogous to $\mathcal{L}_{\text{point}}$ but replaces the point with depth. For camera loss $\mathcal{L}_{\text{cam}}$, we implement a Huber loss $\|\cdot\|_\epsilon$ to supervise the predicted camera $\boldsymbol{E}_i$:

$$\mathcal{L}_{\text{cam}} = \Sigma_{i=1}^{N} \|\boldsymbol{E}_i - \hat{\boldsymbol{E}}_i\|_\epsilon. \tag{7}$$

To supervise the predicted surface normals $\hat{\boldsymbol{E}}_i$, we use Angle Loss (AL), which effectively measures the directional deviation between predicted and ground truth normal vectors. The normal loss function is specifically defined as:

$$\mathcal{L}_{\text{normal}} = \Sigma_{i=1}^{N} \alpha_l \cdot (1 - |\hat{\boldsymbol{N}}_i \cdot \boldsymbol{N}_i|). \tag{8}$$

To enhance robustness in novel views, at each training iteration, we partition the input views I into K candidate context and novel view splits. The pixel overlap rate between the ground truth depth map and camera parameters is computed for each novel view in the context of the candidate context views. The split with the highest pixel overlap rate is selected, with the corresponding context views and novel views being used for further training. Next, based on the selected context images, we regress the 3DGS positions and properties, and render both context view images and novel view images $\hat{I}$. Then, the RGB rendering loss across all views is defined as follows:

$$\mathcal{L}_{rgb} = \Sigma_{i=1}^{N} \|I_i[M_i] - \hat{I}_i[M_i]\| + \lambda_{\text{lpips}}\texttt{LPIPS}(I_i[M_i], \hat{I}_i[M_i]), \tag{9}$$

where M denotes the mask indicating whether the pixels in the current view are visible from the context views, analogous to the novel view mask introduced in [46].

To explicitly supervise the locations of the 3D Gaussian splats, we introduce the depth supervision loss $\mathcal{L}_{\text{gsdepth}}$, which enforces consistency between the ground truth depth map and the depth map predicted by the GS head. The formulation of $\mathcal{L}_{gsdepth}$ follows the same definition as Eq. 6. It is worth noting that, instead of using the depth estimated by the depth head to compute the Gaussian positions, we rely on the GS head to directly predict both the positions and other attributes of the splats. This design choice is further validated in our ablation studies (see Tab. 8). However, due to inherent ambiguities in multi-view rendering and potential noise in the ground truth depth, relying solely on $\mathcal{L}_{\text{rgb}}$ and $\mathcal{L}_{gsdepth}$ often leads to the presence of floating points in the predicted 3DGS. To mitigate this issue, we introduce a gradient consistency loss $\mathcal{L}_{\text{consis}}$, which regularizes the gradients of the GS-rendered depth map $\tilde{D}$ to be consistent with the pseudo depth $\hat{D}$ predicted by the depth head:

$$\mathcal{L}_{\text{consis}} = \Sigma_{i=1}^{N} \|\nabla\hat{D}_i[\hat{M}_i] - \nabla\tilde{D}_i[\hat{M}_i]\|, \tag{10}$$

where $\hat{M}$ is the depth confidence mask corresponding to the top 30%-quantile of the confidence map. Finally, the 3DGS loss is defined as $\mathcal{L}_{\text{3dgs}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{gsdepth}}\mathcal{L}_{\text{gsdepth}} + \lambda_{\text{consis}}\mathcal{L}_{\text{consis}}$.

Figure 3: **Feed-Forward 3D Gaussians Predicted by *WorldMirror* with In-The-Wild Inputs.** Besides real photos, our method generalizes well to AI-created videos spanning diverse styles.

3 Model Evaluation

In this section, we evaluate our approach across four tasks (Sec. 3.2): point map reconstruction, camera pose estimation, surface normal estimation, and novel view synthesis. We also evaluate the effectiveness of different configurations of input priors with a prior-guidance benchmark (Sec. 3.3), and conduct an ablation study to evaluate our design choices (Sec. 3.4). To demonstrate the generalization ability of our method with in-the-wild inputs, we predict the 3D Gaussians (Fig. 8) and point clouds (Fig. 10) with diverse styles of AI-created videos.

3.1 Training Settings

Implementation Details. Our model undergoes a two-phase training process. Initially, we train for 100 epochs using multi-modal prior prompting with a normal head, followed by 50 epochs of fine-tuning with a Gaussian head. Throughout both phases, we implement dynamic image resolutions, maintaining total pixel counts between 100,000 and 250,000, while sampling aspect ratios from 0.5 to 2.0. We employ a dynamic batch sizing approach similar to VGGT, processing 24 images per GPU across a cluster of 32 H20 GPUs. Our optimization strategy features parameter-specific learning rates: 2e-5 for patch embedding layers, 1e-4 for alternated attention modules and pre-trained pointmap, depth, and camera head, and 2e-4 for newly introduced parameters. We use a CosineAnnealing scheduler that gradually decreases from maximum to minimum values following a cosine curve. For our composite loss function, we carefully balance component weights as follows: $\lambda_{\text{points}} = 1.0$, $\lambda_{\text{depth}} = 1.0$, $\lambda_{\text{cam}} = 5.0$, $\lambda_{\text{normal}} = 1.0$, $\lambda_{\text{3dgs}} = 1.0$, $\lambda_{\text{lpips}} = 0.05$, $\lambda_{\text{gsdepth}} = 0.1$, $\lambda_{\text{consis}} = 0.1$.

Dynamic Prior Injection Scheme. Specifically, we randomly toggle each prior modality with a probability of 0.5 during training. When a particular prior is disabled, we set the corresponding tokens to zero. This straightforward approach offers several advantages: it enhances model robustness by forcing the network to handle missing information, enables graceful degradation when certain priors are unavailable during inference, and creates a single unified model capable of operating across different prior combinations.

Curriculum Learning Strategy. During training, we employ a systematic curriculum learning strategy designed to optimize training efficiency and enhance performance by progressing from simple to complex across task sequencing, data scheduling, and resolution.

For task sequencing, initially, we jointly train the multi-modal prior prompting module with other parameters initialized from the pretrained weights of VGGT, which establishes a foundational capability of prior-aware prediction. We then incorporate the normal prediction task into the joint training scheme. Finally, we freeze all model parameters and exclusively train the 3DGS head for 3DGS attributes prediction. This progressive task sequencing strategy ensures effective training for universal geometric prediction with any prior combination.

Table 1: **Point map Reconstruction on 7-Scenes, NRGBD, and DTU.** We report the performance of WorldMirror under different input configurations. The best results are **bold**.

Method	7-Scenes (scene)				NRGBD (scene)				DTU (object)			
	Acc. ↓		Comp. ↓		Acc. ↓		Comp. ↓		Acc. ↓		Comp. ↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Fast3R [60]	0.096	0.065	0.145	0.093	0.135	0.091	0.163	0.104	3.340	1.919	2.929	1.125
CUT3R [53]	0.094	0.051	0.101	0.050	0.104	0.041	0.079	0.031	4.742	2.600	3.400	1.316
FLARE [69]	0.085	0.058	0.142	0.104	0.053	0.024	0.051	0.025	2.541	1.468	3.174	1.420
VGGT [50]	0.046	0.026	0.057	0.034	0.051	0.029	0.066	0.038	1.338	0.779	1.896	0.992
π^3[56]	0.048	0.028	0.072	0.047	0.026	0.015	0.028	0.014	1.198	0.646	1.849	0.607
WorldMirror	0.043	0.026	0.049	0.028	0.041	0.020	0.045	0.019	1.017	0.564	1.780	0.690
WorldMirror (w/ intrinsics)	0.042	0.028	0.048	0.026	0.041	0.020	0.045	0.019	0.977	0.542	1.762	0.682
WorldMirror (w/ depth)	0.038	0.024	0.039	0.023	0.032	0.015	0.031	0.014	0.831	0.506	1.022	0.599
WorldMirror (w/ camera pose)	0.023	0.014	0.036	0.019	0.029	0.018	0.032	0.017	0.990	0.548	1.847	0.686
WorldMirror (w/ intrinsics/depth/camera pose)	**0.018**	**0.011**	**0.023**	**0.014**	**0.016**	**0.011**	**0.014**	**0.010**	**0.735**	**0.461**	**0.935**	**0.550**

Table 2: **Camera Pose Estimation on RealEstate10K, Sintel, and TUM-dynamics.** All datasets are excluded from the training set, except that RealEstate10K was included for CUT3R training.

Method	RealEstate10K (mixed, static)			Sintel (outdoor, dynamic)			TUM-dynamics (indoor, dynamic)		
	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	ATE↓	RPE trans↓	RPE rot↓	ATE↓	RPE trans↓	RPE rot↓
Fast3R[60]	99.05	81.86	61.68	0.371	0.298	13.75	0.090	0.101	1.425
CUT3R [53]	99.82	95.10	81.47	0.217	0.070	0.636	0.047	0.015	0.451
FLARE [69]	99.69	95.23	80.01	0.207	0.090	3.015	0.026	0.013	0.475
VGGT [50]	99.97	93.13	77.62	0.167	0.062	0.491	0.012	0.010	0.312
π^3 [56]	**99.99**	95.62	85.90	**0.074**	**0.040**	**0.282**	0.014	**0.009**	0.312
WorldMirror	**99.99**	**95.81**	**86.28**	0.096	0.058	0.490	**0.010**	**0.009**	**0.297**

For data scheduling, we equip the initial training phase with a comprehensive dataset of both real and synthetic data, which exposes the model to a diverse data distribution for improving the generalization capabilities and preventing overfitting. Following this, the model undergoes a fine-tuning stage using only synthetic data with high-quality annotations of camera, depth, and surface normal, which mitigates the impact of annotation noise inherent in real-world datasets, guiding the model to learn more precise and reliable patterns.

For training resolution, we use a progressive resolution warm-up, beginning with low-resolution inputs and outputs to ensure stable and rapid initial convergence, then gradually increasing the resolution to enhance the model's ability to perceive fine details.

Training Data. The training data comprises a diverse collection of 15 datasets spanning various scene types and capture conditions. This heterogeneous mix includes both established benchmarks and recent collections: DL3DV [30], BlenderMVS [62], TartanAir [55], ASE [35], Unreal4K [49], Habitat [40], MapFree [1], MVS-Synth [17], ArkitScenes [5], ScanNet++ [66], MegaDepth [29], Hypersim [39], Matterport3D [7], Co3dv2 [38], and WildRGBD [57] datasets. This extensive dataset aggregation provides rich supervision across indoor/outdoor environments, real/synthetic scenes, and static/dynamic objects, enabling our model to learn generalizable geometric representations.

3.2 Evaluation on Different Tasks

Point Map Reconstruction. We assess point map reconstruction quality across both scene-level and object-level datasets: 7-Scenes [44], NRGBD [2], and DTU [20]. We use multi-view images with fixed sequence-id mappings from [56] for fair comparison, reporting Accuracy (Acc.) and Completion (Comp.) metrics in Tab. 1. Our method without any priors already surpasses previous SOTA approaches VGGT and π^3, with significant improvements of 10.4% and 17.8% in mean accuracy on 7-Scenes and DTU, respectively. Incorporating a single prior can further enhance performance, while the combination of all priors achieves optimal results, which delivers clear gains of 58.1% and 53.1% in mean accuracy on 7-Scenes and NRGBD compared to our no-prior baseline. These results clearly demonstrate our model's ability to effectively leverage prior information for better reconstruction.

Camera Pose Estimation. Following the protocol of [56], we test camera pose estimation on three unseen datasets: RealEstate10K [70], Sintel [6], and TUM-dynamics [47]. For RealEstate10K, we select 10 fixed images per sequence and examine all pairwise combinations, measuring Relative

Table 3: **Monocular and Video Depth Estimation on NYUv2, Sintel, and KITTI.**

Method	**NYU-v2** (Monocular)		**Sintel** (Monocular)		**KITTI** (Video)		**Sintel** (Video)	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
DUSt3R [54]	0.081	0.909	0.488	0.532	0.143	0.814	0.662	0.434
MASt3R [28]	0.11	0.865	0.413	0.569	0.115	0.848	0.558	0.487
MonST3R [68]	0.094	0.887	0.492	0.525	0.107	0.884	0.399	0.519
Fast3R [60]	0.093	0.898	0.544	0.509	0.138	0.834	0.638	0.422
CUT3R [53]	0.081	0.914	0.418	0.52	0.122	0.876	0.417	0.507
FLARE [69]	0.089	0.898	0.606	0.402	0.356	0.57	0.729	0.336
VGGT [50]	0.056	0.951	0.606	0.599	0.062	0.969	0.299	0.638
π^3 [56]	0.054	0.956	**0.277**	0.614	**0.038**	**0.986**	**0.233**	0.664
WorldMirror	**0.052**	**0.957**	0.339	**0.624**	0.063	0.968	0.289	**0.668**

Table 4: **Surface Normal Estimation on ScanNet, NYUv2, and iBims-1.** We compare with both regression-based and diffusion-based surface normal estimation approaches. EESNU is trained on ScanNet, thus its in-domain performance is omitted.

Method	**ScanNet**				**NYUv2**				**iBims-1**			
	mean ↓	med ↓	22.5° ↑	30° ↑	mean ↓	med ↓	22.5° ↑	30° ↑	mean ↓	med ↓	22.5° ↑	30° ↑
OASIS [9]	32.8	28.5	38.5	52.6	29.2	23.4	48.4	60.7	32.6	24.6	46.6	57.4
EESNU [3]	-	-	-	-	16.2	8.5	77.2	83.5	20.0	8.4	73.4	78.2
Omnidata v1 [13]	22.9	12.3	66.1	73.2	23.1	12.9	66.3	73.6	19.0	7.5	76.1	80.1
Omnidata v2 [23]	16.2	8.5	79.5	84.7	17.2	9.7	76.5	83.0	18.2	7.0	77.4	81.1
DSine [4]	16.2	8.3	78.7	84.4	16.4	8.4	77.7	83.5	17.1	**6.1**	79.0	82.3
GeoWizard [15]	16.7	9.5	78.3	84.2	19.5	11.7	74.5	81.6	20.4	9.4	76.4	80.6
StableNormal [64]	16.0	9.9	81.5	86.5	18.5	11.2	77.5	83.6	17.9	8.5	**80.4**	**83.9**
WorldMirror	**13.8**	**7.3**	**82.5**	**87.3**	**15.1**	**8.0**	**80.1**	**85.7**	**16.6**	6.4	80.1	83.7

Table 5: **Novel View Synthesis on RealEstate10K and DL3DV.** We compare with feed-forward 3DGS methods under sparse and dense-view settings. FLARE focuses on sparse views NVS and thus its performance under dense-view settings is omitted.

Method	**RealEstate10K** (2 views)			**DL3DV** (8 views)			**RealEstate10K** (32 views)			**DL3DV** (64 views)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FLARE [69]	16.33	0.574	0.410	15.35	0.516	0.591	-	-	-	-	-	-
AnySplat [22]	17.62	0.616	0.242	18.31	0.569	0.258	19.96	0.718	0.234	18.40	0.602	0.286
WorldMirror	20.62	0.706	0.187	20.92	0.667	0.203	25.14	0.859	0.109	21.25	0.703	0.223
WorldMirror (w/ intrinsics)	22.03	0.765	0.165	22.08	0.723	0.175	25.71	0.877	0.101	21.55	0.731	0.207
WorldMirror (w/ camera pose)	20.84	0.713	0.182	21.18	0.674	0.197	25.14	0.865	0.107	21.28	0.700	0.222
WorldMirror (w/ intrinsics/camera pose)	**22.30**	**0.774**	**0.155**	**22.15**	**0.726**	**0.174**	**25.77**	**0.879**	**0.101**	**21.66**	**0.736**	**0.204**

Rotation Accuracy (RRA), Relative Translation Accuracy (RTA), and Area Under the Curve (AUC) at a 30-degree threshold. For Sintel and TUM-dynamics, we report Absolute Trajectory Error (ATE), Relative Pose Error for translation (RPE trans), and rotation (RPE rot). Tab. 2 demonstrates strong results: our method achieves superior zero-shot performance on RealEstate10K and TUM-dynamics, while maintaining competitive results on Sintel. The performance on Sintel, though slightly below the best methods, is reasonable given the limited outdoor dynamic scenes in our training data.

Monocular and Video Depth Estimation In Table 3, we evaluate *WorldMirror* in comparison with contemporary approaches for both single-view and sequential depth estimation across diverse input scenarios. Despite *WorldMirror* not being explicitly optimized for monocular metric depth inference, it delivers performance that matches or exceeds current leading methods. When processing video sequences, *WorldMirror* produces results that rival specialized feed-forward reconstruction frameworks. We note a modest performance gap on the KITTI benchmark relative to π^3, which we attribute to the under-representation of urban driving environments in our training distribution. Future iterations of our work will incorporate a more comprehensive collection of street-level imagery to enhance generalization to such scenarios.

Surface Normal Estimation. Following the protocol from [4], we evaluate surface normal estimation on three datasets: iBims-1[27], NYUv2 [45], and ScanNet [11]. We measure angular error between predicted and ground truth normal maps, reporting both mean and median errors along with the percentage of pixels below error thresholds of 22.5° and 30.0°. Tab. 4 presents our method's performance across three datasets, demonstrating substantial improvements over existing approaches.

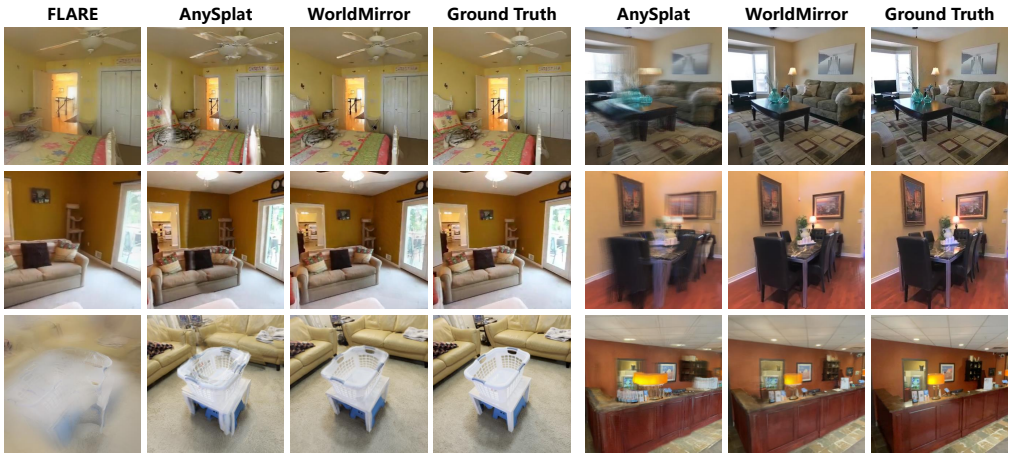

Figure 4: **Qualitative Comparisons of Novel View Synthesis.** We compare with FLARE and AnySplat on RealEstate10K and DL3DV. The first four columns correspond to the sparse-view setting, while the latter three correspond to the dense-view setting. Our approach surpasses baselines in both appearance fidelity and geometric perception.

Table 6: **Novel View Synthesis with 3DGS Optimization on RealEsate10K, DL3DV, and VRNeRF.** In Post-Optimization, the *random point cloud* refers to initializing Gaussian positions randomly, whereas the *predicted point cloud* uses the point cloud estimated by our method as the initialization of Gaussian positions.

Method	Iterations	RealEstate10K (32 views)				DL3DV (64 views)				VRNeRF (64 views)			
		PSNR ↑	SSIM ↑	LPIPS ↓	Time ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Time ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Time ↓
Feedforward													
AnySplat	-	19.96	0.718	0.234	<2s	18.40	0.602	0.286	<2s	22.11	0.759	0.288	<2s
WorldMirror	-	25.14	0.859	0.109	<2s	21.25	0.703	0.223	<2s	25.77	0.830	**0.208**	<2s
Post Optimization													
random points cloud	3,000	26.03	0.875	0.145	19s	23.61	0.765	0.244	21s	**26.45**	0.840	0.259	21s
predicted points cloud	1,000	27.29	0.906	0.092	10s	23.43	0.772	0.248	12s	25.19	0.841	0.257	11s
AnySplat	1,000	23.85	0.834	0.192	23s	20.84	0.695	0.287	55s	23.19	0.782	0.322	33s
AnySplat	3,000	26.03	0.870	0.155	56s	22.20	0.723	0.226	126s	24.64	0.798	0.272	65s
WorldMirror	1,000	**27.79**	**0.915**	**0.076**	23s	**23.86**	**0.786**	**0.172**	45s	25.98	**0.845**	0.214	38s

The consistent gains across diverse datasets indicate that multi-task frameworks leveraging shared representations can effectively outperform specialized single-task methods.

Novel View Synthesis. We evaluate zero-shot novel view synthesis on three datasets: RealEstate10K [70], DL3DV [30], and VR-NeRF [59] under both sparse-view and dense-view settings. For RealEstate10K, we randomly sample 200 scenes from the NopoSplat [63] test split, using 3 novel views per scene in the sparse-view setting and 4 novel views per scene in the dense-view setting. For DL3DV, we follow the FLARE test split and evaluate in 112 unseen scenes, each containing 9 novel views. For VR-NeRF, consistent with AnySplat, we select 5 scenes, each with 64 input views and 6 novel views. For calculating the rendering metrics, we follow the *test-time camera pose alignment* introduced by AnySplat to ensure fair evaluation. Tab. 5 reports the quantitative evaluation results for novel view synthesis under the feed-forward setting. Our method achieves substantial improvements over the previous state-of-the-art AnySplat, with consistent gains across all metrics on both datasets, demonstrating the effectiveness of our unified geometric representation for high-quality view synthesis.

Novel View Synthesis with Optimization. Although recent feed-forward pipelines are capable of synthesizing competitive 3D Gaussian splats (3DGS) within seconds, they inevitably suffer from errors introduced by single-pass predictions, such as suboptimal Gaussian placement and appearance. We hypothesize that incorporating a brief post-optimization stage—initialized with either our predicted point cloud or 3DGS primitives—can significantly refine both geometry and appearance at only modest additional cost, thereby accelerating the convergence of 3DGS training and enhancing rendering quality. As shown in Tab. 6, we compare (i) feed-forward baselines and (ii) post-optimization with 3,000 or 1,000 iterations, initialized either from a random point cloud or from

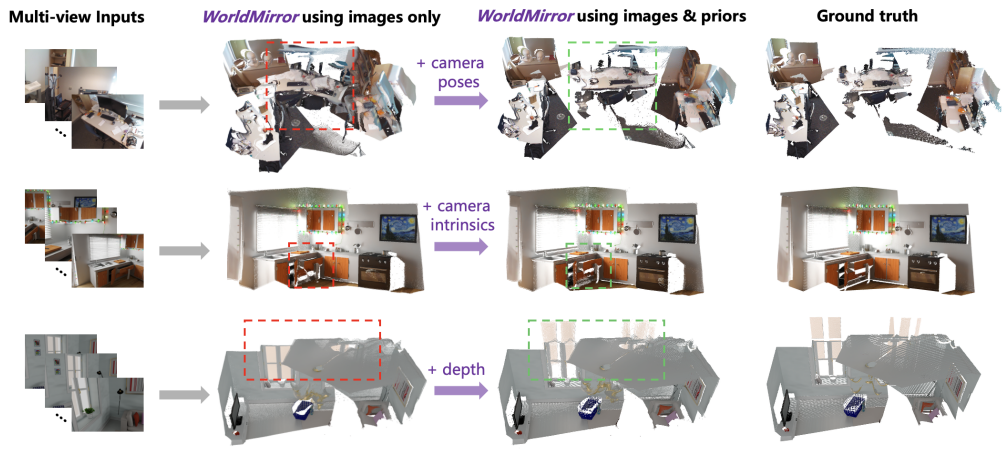

Figure 5: **Geometric Priors Unlock Enhanced Scene Reconstruction of *WorldMirror*.** (`Top`) Camera poses help the model to capture relative view positions accurately. (`Middle`) Calibrated intrinsic enhances the reconstruction by enabling precise projection modeling and geometry alignment. (`Bottom`) Depth guidance enables the network to better handle challenging reconstruction scenarios, like perspective distortion, unusual geometric configurations, or partial occlusions.

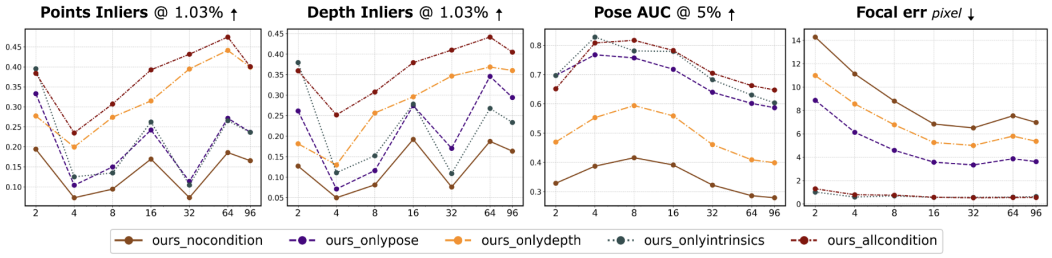

Figure 6: **Geometric Priors Boosts Model's Feed-Forward Performance across All Tasks.** Incorporating a single modality not only enhances predictions for its corresponding task but also improves performance across other tasks. This suggests that modal information enables the model to develop a more comprehensive understanding of the overall geometry.

feed-forward 3DGS primitives. The camera parameters for optimizing 3DGS are obtained from the feed-forward outputs of the chosen method. Our predicted point cloud, camera, and 3DGS primitives provide a robust and high-quality initialization for 3DGS optimization, significantly accelerating the training process and consistently surpassing baseline methods across all metrics.

3.3 Evaluation on Different Input Configurations

To demonstrate the benefits of incorporating priors into model predictions, we evaluate model performance across various input configurations. We present four key metrics: the inlier ratio at a relative threshold of 1.03% of points and depths, the area under the curve at a 5° error threshold (AUC@5), and the average focal error in pixels, measured across the ETH3D [43] and DTU [20] datasets. As shown in Fig.6, incorporating even a single modality prior yields dual benefits: it enhances both the corresponding task prediction and the model's capacity to infer other geometric attributes. Fig.5 illustrates how different priors contribute to reconstruction quality. Camera poses enable the model to capture global scene geometry, calibrated intrinsics resolve scale ambiguity, while depth priors offer pixel-level constraints that prove particularly valuable for reconstructing geometrically complex regions. These findings confirm that multi-modal priors work synergistically, where each modality provides complementary geometric constraints that collectively improve the model's understanding of 3D scene structure.

Table 7: **Prior Embedding Ablation.** Results are averaged over ETH3D and DTU datasets with 10 views as input. 'Single token' offers both superior performance and high efficiency.

Prior embedding	Extra Params	Focal acc@1.03↑	Depth τ@1.03↑	RRA@5↑	Pose RTA@5↑	AUC@5↑	Point τ@1.03↑	Avg.↑
Input: images & poses								
Dense Plücker	9.02M	33.07	31.00	98.59	93.52	72.74	33.74	60.44
Single Token	1.06M	33.82	28.02	98.89	92.57	74.55	38.51	61.06
Input: images & intrinsics								
Dense Raymap	6.65M	86.48	29.36	97.17	88.48	60.57	37.40	66.58
Single Token	1.06M	84.43	34.70	98.18	93.64	66.52	36.29	68.96

Table 8: **Novel View Synthsis Ablation.** Results are from RealEstate10K, DL3DV, and VR-NeRF.

Method	**RealEstate10K** (2 views)			**DL3DV** (8 views)			**VR-NeRF** (32 views)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o GT Cameras	**20.30**	0.691	0.193	20.69	0.666	0.206	24.76	0.788	0.197
w/o Novel Views	18.51	0.651	0.215	20.21	0.664	**0.196**	24.35	0.781	0.199
w/o GS DPT	20.28	0.693	0.195	20.55	0.667	0.218	25.08	0.798	**0.191**
Ours	20.29	**0.693**	**0.192**	**20.91**	**0.671**	0.198	**25.75**	**0.811**	0.198

Figure 7: ***WorldMirror*** **Improves Surface Reconstruction with Predicted Normal Maps.**

3.4 Ablation Study

Prior Embedding Ablation. We explore different ways of embedding priors in Tab. 7. For camera poses, we experiment with (1) dense Plücker ray embeddings that are added element-wise to the image tokens, and (2) a single token concatenation approach where the pose is compressed into a single token and concatenated to the sequence. For camera intrinsics, we similarly compare dense raymap embeddings that are added to the image tokens versus a single token. Our experiments reveal that the single token approach achieves better performance for embedding both camera poses and intrinsics, suggesting that a compact global representation is more effective than dense per-pixel conditioning while being more efficient.

Novel View Synthesis Ablation. Tab. 8 reports ablation analysis on the novel view synthesis: (1) To examine the importance of using ground-truth camera parameters for novel view rendering, we replace the ground-truth poses and intrinsic matrices in our method with those predicted by the camera head for computing 3DGS positions and rendering. (2) To assess the necessity of supervising 3DGS rendering not only on input views but also on novel views, we perform an ablation similar to [22], where no novel-view rendering loss is applied. (3) The GS head predicts all Gaussian attributes except positions, while the positions are derived from the depth maps estimated by the Depth head. These studies confirm that both our 3DGS prediction framework and training strategy are crucial, and removing any component degrades novel view rendering performance.

3.5 Applications

Surface Reconstruction. *WorldMirror* supports high-quality 3D surface reconstruction with the predicted smooth normal maps. As shown in Fig. 7, by leveraging the predicted normals instead of traditional geometric normal estimation from point clouds, *WorldMirror* produces a cleaner surface with sharp details via Poisson surface reconstruction [25].

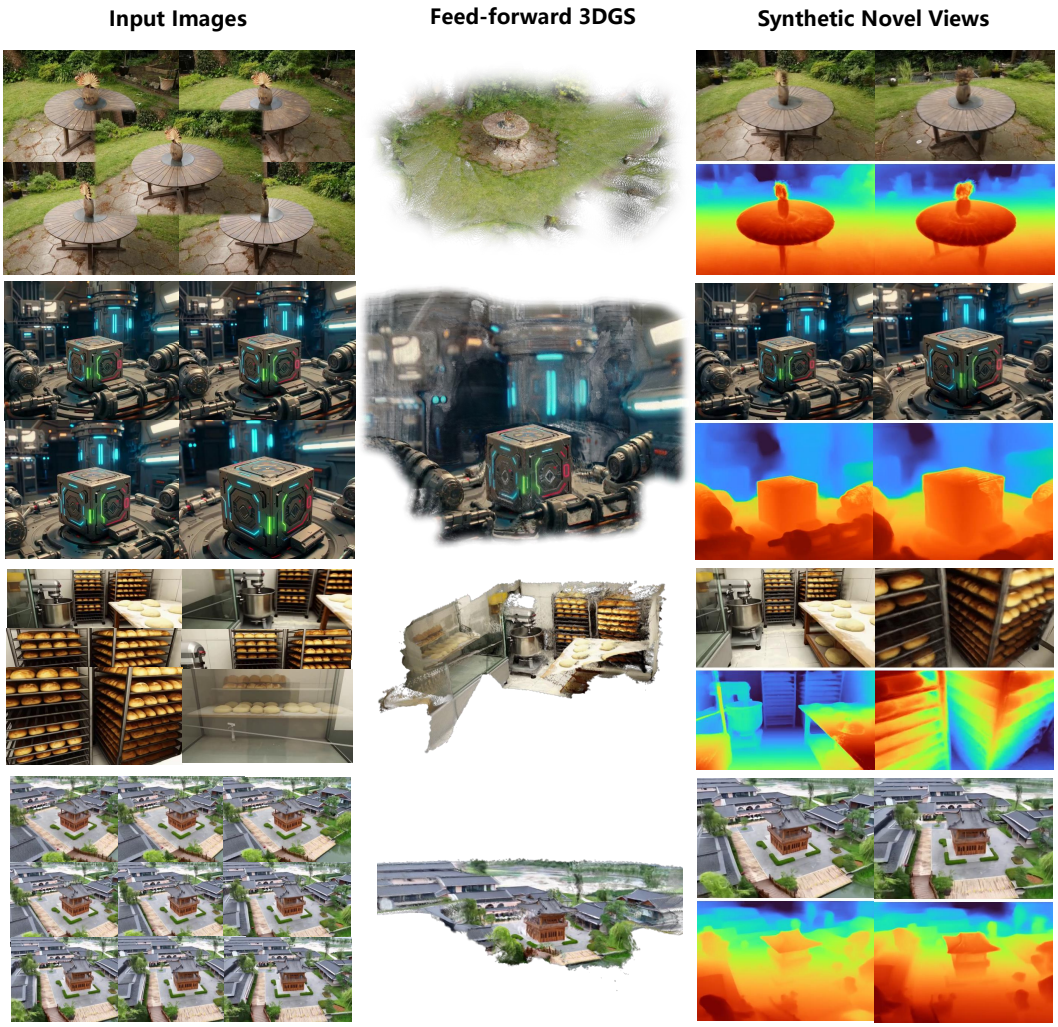

Figure 8: **Visual Results of Feed-Forward 3D Gaussians Generated by *WorldMirror*.**

3.6 More Visual Results

Novel View Synthesis. In Fig. 8, we present additional results of feedforward Gaussians and their corresponding novel view renderings. Whether the input consists of AI-generated videos or real multi-view images, our method consistently infers 3D Gaussian splatting with plausible geometric structures and renders high-quality novel view images. This demonstrates that our model generalizes effectively across diverse input scenarios.

Point Map Reconstruction. We provide additional visual comparisons of point map reconstruction in Fig. 9 and Fig. 10. Fig. 9 features selected scenes from 7-scenes, NRGBD, and DTU datasets, where comparisons with ground truth reveal that *WorldMirror* produces more consistent reconstructions, particularly when processing sparse viewpoints that require inference of spatial distributions. In Fig. 10, we evaluate model performance on in-the-wild images by processing both video generation model outputs and real-world multi-view captures. The results demonstrate that *WorldMirror* generates geometrically coherent and plausible reconstructions across these diverse inputs, highlighting its strong generalization capabilities.

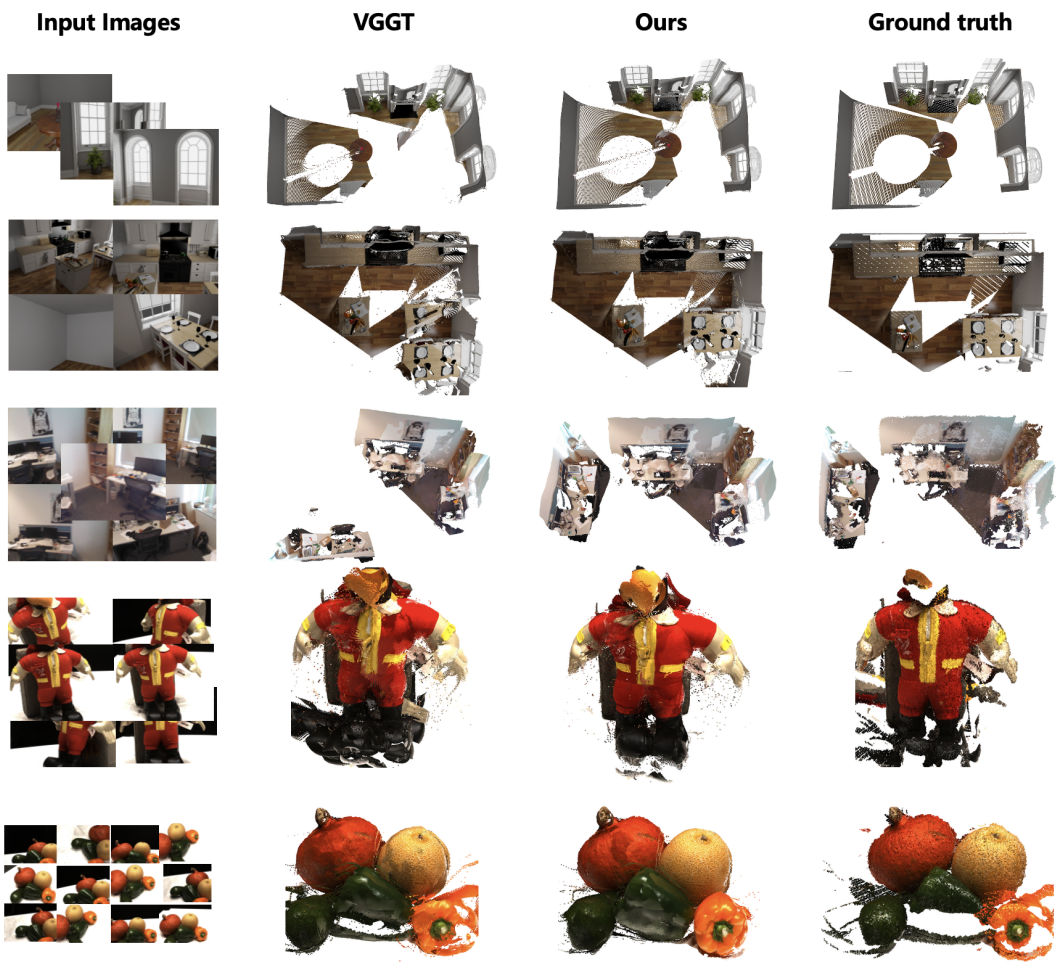

Figure 9: **Visual Comparisons on 7-Scenes, NRGBD, and DTU datasets.** *WorldMirror* delivers superior reconstruction fidelity compared to VGGT, effectively capturing spatial relationships within scenes while producing geometrically coherent structures.

4 Related Work

4.1 Feed-Forward 3D Reconstruction.

Feed-forward 3D reconstruction models have recently emerged as powerful alternatives to traditional SfM/MVS pipelines by directly regressing 3D structure. DUSt3R [54] pioneers this direction with point map prediction, while Fast3R [60] improves its scalability. VGGT [50] further introduces large-scale multi-task learning, with subsequent variants that remove reference-view bias [56] and extend to kilometer-scale sequences [12]. Meanwhile, Dens3R [14] introduces a dense prediction backbone for joint estimation of geometric attributes. Building on these advances, *WorldMirror* unifies an even broader range of 3D tasks, including camera poses, depth, surface normals, point maps, and novel view synthesis, in one feed-forward pass.

4.2 3D Prior Guidance.

Traditional optimization-based methods like COLMAP [42] incorporate known camera parameters to improve reconstruction quality. Recent learning-based approaches have also explored different forms of guidance: UniDepth [36] optionally uses camera intrinsics for improved monocular depth estimation, while some video diffusion models [16, 18, 48] demonstrate how camera trajectories can guide consistent content generation. More recently, Pow3R [19] extends DUSt3R [54] with

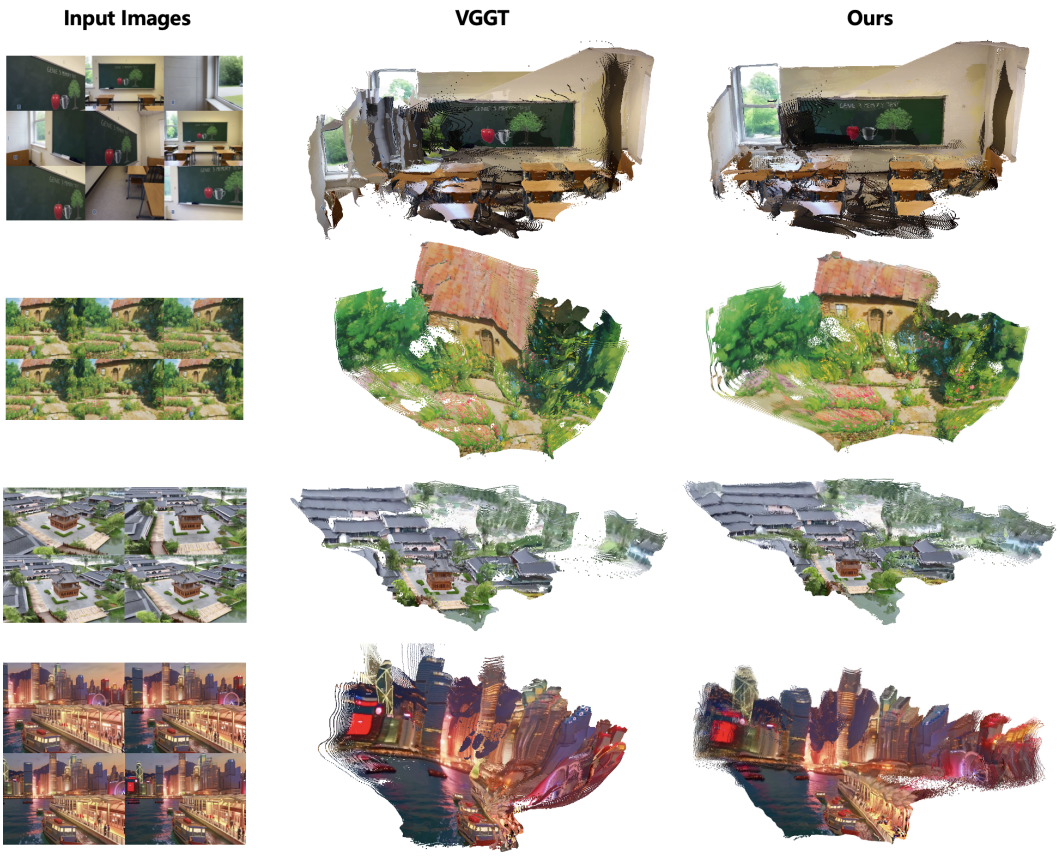

Figure 10: **Visual Comparisons of In-The-Wild Multi-View 3D Reconstruction.** *WorldMirror* delivers superior reconstruction fidelity with in-the-wild images as input, generating more plausible results in challenging scenarios compared to VGGT. Our approach effectively resolves complex spatial arrangements and maintains geometric consistency even when confronted with difficult viewing conditions, occlusions, or intricate environmental structures.

additional modalities as input but remains limited to sparse-view inputs within the "3R" paradigms. The integration of more modalities into dense regression frameworks like VGGT remains unexplored. In this paper, we present the first systematic exploration of multi-modal geometric prior injection within dense multi-view reconstruction frameworks.

4.3 Generalizable Novel View Synthesis.

Novel view synthesis (NVS) has been extensively studied with representations such as NeRF [33] and 3D Gaussian Splatting [26], which achieve photorealistic results but typically require dense-view training for each scene. Early generalizable NVS methods [67, 8, 58, 31] take sparse-view images with known intrinsics and poses as input to produce 3D scenes or novel views. While effective for sparse inputs, these approaches depend on accurate calibration or fixed view counts [10, 34]. Pose-free methods [21, 52, 63] instead pursue end-to-end reconstruction directly from images. FLARE [69] introduces a cascaded pose-geometry-appearance pipeline, while AnySplat [22] combines 3D foundation models with 3D Gaussians for real-time NVS from uncalibrated images. We advance beyond these methods by enabling pose-free novel view synthesis with flexible input view counts, optional prior incorporation, and superior rendering quality.

5 Conclusion

We presented *WorldMirror*, a unified feed-forward model that addresses versatile 3D reconstruction tasks. By flexibly incorporating diverse geometric priors and generating multiple 3D representations simultaneously, our framework demonstrates that a single model can effectively handle various 3D reconstruction tasks without task-specific specialization. *WorldMirror* achieves state-of-the-art performance across dense reconstruction, multi-view depth estimation, surface normal prediction, and novel view synthesis, while maintaining feed-forward efficiency. The model's ability to leverage available priors enables robust reconstruction in challenging scenarios, and its multi-task design ensures geometric consistency across different outputs. Our work shows that unified, prior-aware architectures offer a promising direction for comprehensive and efficient 3D scene understanding.

Limitations and Future Works. Despite the promising results achieved by our approach, several limitations remain. First, our method demonstrates suboptimal performance on dynamic scenes and autonomous driving environments, primarily due to the under-representation of such data in our training distribution. We plan to address this through strategic dataset expansion to enhance model generalization. Additionally, our current implementation supports input resolutions ranging from 300 to 700 pixels and cannot effectively handle scenarios where the number of input views reaches into the thousands. This constraint becomes particularly apparent when running on consumer-grade GPUs. Future work will explore computational optimizations to improve model efficiency and enable processing of longer visual sequences with reduced memory requirements.

Contributors

- **Project Sponsors:** Jie Jiang, Linus, Yuhong Liu, Peng Chen
- **Project Leaders:** Chunchao Guo, Tengfei Wang
- **Core Contributors:** Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang
- **Contributors:** Xuhui Zuo, Chenjie Cao, Haoyuan Wang, Lifu Wang, Yulin Tsai, Yonghao Tan, Chao Zhang, Hao Zhang, Runzhou Wu, Yifu Sun, Lin Niu, Xiang Yuan

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022.

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022.

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021.

[4] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024.

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[6] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021.

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[8] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024.

[9] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 679–688, 2020.

[10] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024.

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[12] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025.

[13] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.

[14] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, et al. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025.

[15] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

[16] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

[17] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025.

[19] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1071–1081, 2025.

[20] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

[21] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. *arXiv preprint arXiv:2310.01410*, 2023.

[22] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025.

[23] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022.

[24] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pages 18–35. Springer, 2024.

[25] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

[26] 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.

[27] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[28] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[29] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[31] Yifan Liu, Keyu Fan, Weihao Yu, Chenxin Li, Hao Lu, and Yixuan Yuan. Monosplat: Generalizable 3d gaussian splatting from monocular depth foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21570–21579, 2025.

[32] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

[33] 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.

[34] Zhiyuan Min, Yawei Luo, Jianwen Sun, and Yi Yang. Epipolar-free 3d gaussian splatting for generalizable novel view synthesis. *Advances in Neural Information Processing Systems*, 37:39573–39596, 2024.

[35] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023.

[36] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024.

[37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[38] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021.

[39] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

[40] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[42] 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*, pages 501–518. Springer, 2016.

[43] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[44] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013.

[45] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[46] 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.

[47] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.

[48] HunyuanWorld Team Tencent. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels, 2025.

[49] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8942–8952, 2021.

[50] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025.

[51] 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.

[52] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023.

[53] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[54] 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*, pages 20697–20709, 2024.

[55] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[56] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[57] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024.

[58] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16453–16463, 2025.

[59] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, et al. Vr-nerf: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.

[60] 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*, pages 21924–21935, 2025.

[61] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024.

[62] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020.

[63] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

[64] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.

[65] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025.

[66] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[67] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021.

[68] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[69] 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*, pages 21936–21947, 2025.

[70] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.