

MV-DUSt3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds

Zhenggang Tang^{1,2}, Yuchen Fan¹, Dilin Wang¹, Hongyu Xu¹, Rakesh Ranjan¹, Alexander Schwing², Zhicheng Yan^{1,†}

¹Meta Reality Labs, ²University of Illinois Urbana-Champaign

†project lead

Recent sparse multi-view scene reconstruction advances like DUSt3R and MASt3R no longer require camera calibration and camera pose estimation. However, they only process a pair of views at a time to infer pixel-aligned pointmaps. When dealing with more than two views, a combinatorial number of error prone pairwise reconstructions are usually followed by an expensive global optimization, which often fails to rectify the pairwise reconstruction errors. To handle more views, reduce errors, and improve inference time, we propose the fast single-stage feed-forward network MV-DUSt3R. At its core are multi-view decoder blocks which exchange information across any number of views while considering one reference view. To make our method robust to reference view selection, we further propose MV-DUSt3R+, which employs cross-reference-view blocks to fuse information across different reference view choices. To further enable novel view synthesis, we extend both by adding and jointly training Gaussian splatting heads. Experiments on multi-view stereo reconstruction, multi-view pose estimation, and novel view synthesis confirm that our methods improve significantly upon prior art. Code will be released.

Correspondence: zyan3@meta.com

Website: <https://mv-dust3rp.github.io/>



Figure 1 The proposed Multi-View Dense Unconstrained Stereo 3D Reconstruction Prime (MV-DUSt3R+) is able to reconstructs large scenes from multiple pose-free RGB views. **Top row:** one single-room scene and one large multi-room scene reconstructed by MV-DUSt3R+ in 0.89 and 1.54 seconds using 12 and 20 input views respectively (only a subset is shown for visualization). **Bottom row:** MV-DUSt3R+ is able to synthesize novel views by predicting pixel-aligned Gaussian parameters. Reconstruction of such large scenes are challenging for prior methods (*e.g.* DUSt3R (Wang et al., 2024)). See figure 6 and appendix for more results with comparison.

1 Introduction

Multi-view scene reconstruction as shown in figure 1 has been a fundamental task in 3D computer vision for decades (Hartley and Zisserman, 2003). It is widely applicable in mixed reality (Avetisyan et al., 2024),

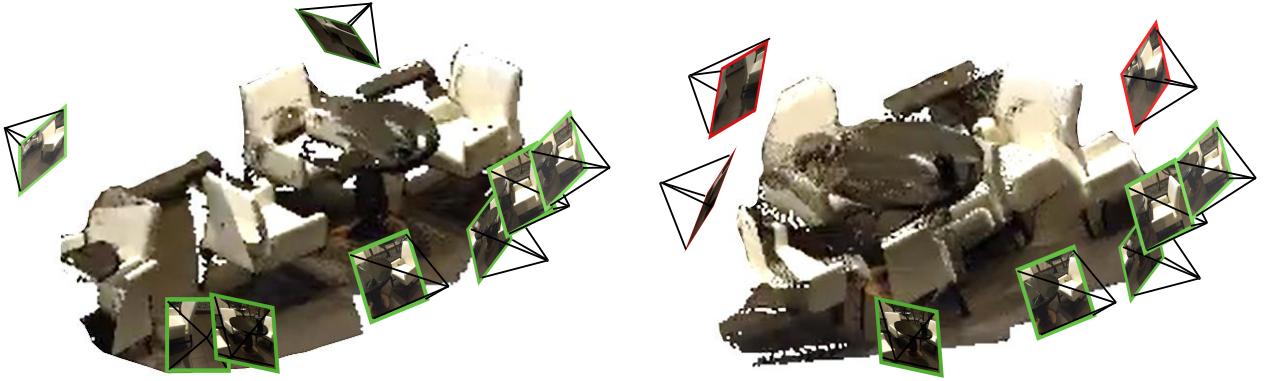


Figure 2 **Left:** Groundtruth scene with 8 views: Three chairs surrounding one table and one more chair next to another table. **Right:** reconstruction and pose estimation of DUST3R with global optimization: all chairs incorrectly surround one table. Wrong poses are marked in red.

city reconstruction (Turki et al., 2022), autonomous driving simulation (Khan et al., 2024; Xiao et al., 2021), robotics (Zeng et al., 2020) and archaeology (Peppa et al., 2018). Classic methods decompose multi-view scene reconstruction into sub-tasks, including camera calibration (Zhang et al., 2011; Bougnoux, 1998), pose estimation (Quan and Lan, 1999; Brachmann et al., 2017), feature detection and matching (Lowe, 2004; Rublee et al., 2011; Yi et al., 2016), structure from motion (SfM) (Schonberger and Frahm, 2016), bundle adjustment (Triggs et al., 2000; Cadena et al., 2016), etc., and assemble individual components into a pipeline. Recent approaches adopt a learning-based paradigm for those sub-tasks (Fu et al., 2022), explore different neural scene representations (*e.g.*, Neural Signed Distance Functions (Park et al., 2019; Ma et al., 2023; Fu et al., 2022), Neural Radiance Fields (Mildenhall et al., 2021; Chan et al., 2022), Gaussian Splatting (Kerbl et al., 2023; Huang et al., 2024)), and build more end-to-end pipelines to reconstruct objects (Tang et al., 2024; Xu et al., 2024) and scenes (Zhang et al., 2024; Charatan et al., 2024; Chen et al., 2024). While these approaches have enjoyed quite some success, they often require prior knowledge or nontrivial pre-processing to obtain camera parameters and poses.

More recently, novel multi-view scene reconstruction approaches, such as DUST3R (Wang et al., 2024) and MAST3R (Leroy et al., 2024), directly process an unordered set of unposed rgb views, *i.e.*, camera intrinsics and poses are unknown. These methods process two views, *i.e.*, a chosen reference view and another source view, at a time, and directly infer pixel-aligned 3D pointmaps in the reference view’s camera coordinate system. To handle a larger set of input views, a combinatorial number of pairwise inferences is followed by a second stage of global optimization to align the local pairwise reconstructions into a single global coordinate system.

While evaluation results of these methods are promising on object-centric DTU data (Aanæs et al., 2016), we point out inefficiencies in reconstructing scenes, as stereo cues in 2-view input could be ambiguous. Further, despite plausible reconstructions of individual 2-view inputs, conflicts often arise when aligning them in a global coordinate system. Such conflicts can be challenging to resolve by a global optimization, which only rotates pairwise predictions but doesn’t rectify wrong pairwise matches. As a consequence, scene reconstructions exhibit misaligned pointmaps. One example is shown in figure 2.

To address the aforementioned issues, we propose the single-stage network **Multi-View Dense Unconstrained Stereo 3D Reconstruction (MV-DUST3R)**, which jointly processes a large number of input views in one feed-forward pass, and completely removes the cascaded global optimization used in prior arts. To achieve this, we employ multi-view decoder blocks, which jointly learn not only all pairwise relationships between a chosen reference view and all the other source views, but also appropriately address pairwise relationships among all source views. Moreover, our training recipe encourages the predicted per-view pointmaps to adhere to the same reference camera coordinate system, which waives the need for a subsequent global optimization.

When reconstructing a large scene from sparse multi-view images, the stereo cues between the one selected reference view and certain source views could be insufficient. This is because significant changes in camera

poses make it difficult to directly infer the relation between the reference view and those source views. Therefore, long-range information propagation is required for those source views. To handle this efficiently, we further present **MV-DUSt3R+**. It operates on *a set of reference views* and employs Cross-Reference-View attention blocks for effective long-range information propagation. See [figure 1](#) for its reconstructions.

On the three benchmark scene-level datasets HM3D ([Ramakrishnan et al., 2021](#)), ScanNet ([Dai et al., 2017](#)), and MP3D ([Chang et al., 2017](#)), we demonstrate that MV-DUSt3R achieves significantly better results on the tasks of Multi-View Stereo (MVS) reconstruction and Multi-View Pose Estimation (MVPE) while being $48 \sim 78\times$ faster than DUSt3R. In addition, MV-DUSt3R+ is able to improve the reconstruction quality especially on harder settings, while still inferring one order of magnitude faster than DUSt3R.

To extend both of our methods towards Novel View Synthesis (NVS), we further attach lightweight prediction heads which regress to 3D Gaussian attributes. The predicted per-view Gaussian primitives are transformed into the coordinate of a target view before splatting-based rendering ([Kerbl et al., 2023](#)). Using this, we also show that our models outperform DUSt3R with heuristically designed 3D Gaussian parameters under the standard photometric evaluation protocol. The gains can be attributed to the more accurate predictions of the Gaussian locations by our method. We summarize our contributions as follows:

- We present **MV-DUSt3R**, a novel feed-forward network for pose-free scene reconstruction from sparse multi-view input. It not only runs $48 \sim 78\times$ faster than DUSt3R for $4 \sim 24$ views, but also reduces Chamfer distance on 3 challenging evaluation datasets HM3D ([Ramakrishnan et al., 2021](#)), ScanNet ([Dai et al., 2017](#)), and MP3D ([Chang et al., 2017](#)) by $2.8\times$, $2\times$ and $1.6\times$ for smaller scenes of average size $2.2, 7.5, 19.3 (m^2)$ with 4-view input, and $3.2\times, 1.9\times$ and $2.1\times$ for larger scenes of average size $3.3, 17.9, 37.3 (m^2)$ with 24-view input.
- We present **MV-DUSt3R+**, which improves MV-DUSt3R by using multiple reference views, addressing the challenges which occur when inferring relations between all input views via a single reference view. We validate, MV-DUSt3R+ performs well across all tasks, number of views, and on all three datasets. For example, for MVS reconstruction, it further reduces Chamfer distance on 3 datasets by $2.6\times, 1.6\times, 1.8\times$ for large scenes with 24-view input, while still running $14\times$ faster than DUSt3R.
- We extend both networks to support NVS by adding Gaussian splatting heads to predict per-pixel Gaussian attributes. With joint training of all layers using both reconstruction loss and view rendering loss, we demonstrate that the model outperforms a DUSt3R-based baseline significantly.

2 Related Work

Structure-from-Motion (SfM). SfM methods reconstruct sparse scene geometry from a set of images and estimate individual camera poses. For this, SfM is often addressed in a few independent steps, including detecting/describing/matching local features across multiple views (*e.g.*, SIFT ([Lowe, 2004](#)), ORB ([Rublee et al., 2011](#)), LIFT ([Yi et al., 2016](#))), triangulating features to estimate sparse 3D geometry and camera poses (*e.g.*, COLMAP ([Schonberger and Frahm, 2016](#))), applying bundle adjustment over many views (see [Triggs et al. \(2000\)](#) for an overview), *etc*. Though steady progress has been made in the past decades ([Özyeşil et al., 2017](#)), and a large number of applications have been enabled ([Westoby et al., 2012](#); [Ighlaut et al., 2019](#); [Carrivick et al., 2016](#)), the classic SfM pipeline solves sub-tasks individually and sequentially, accumulating errors. More recent SfM methods improve traditional pipeline with learnable components ([He et al., 2024](#); [Wang et al., 2023a](#)). MAST3R-SfM ([Duisterhof et al., 2024](#)) extends MAST3R ([Leroy et al., 2024](#)), which only produces local reconstructions for 2-view input, to perform global optimization for aligning local reconstructions via gradient descent to minimize 3D matching loss.

Multi-View Stereo. MVS reconstructs dense 3D scene geometry from multiple views ([Furukawa et al., 2015](#)), often in the form of 3D points. In the classic PatchMatch-based framework ([Zheng et al., 2014](#)), per-pixel depth in the reference image is estimated from a set of unstructured source images via patch matching under a homography transform ([Schönberger et al., 2016](#)). Subsequent work has substantially improved feature matching ([Wang et al., 2023c](#); [Zhou et al., 2018a](#); [Wang et al., 2021](#)) and depth estimation ([Galliani et al., 2015](#); [Schmied et al., 2023](#); [Xu and Tao, 2019](#)). More recent learning-based approaches ([Yao et al., 2018](#); [Wang and Shen, 2018](#)) often build an end-to-end pipeline, where deep models extract visual features, model

cross-view correspondences (*e.g.*, cost volume (Gu et al., 2020)), and regress depth maps (Yao et al., 2019). Note, with few exceptions (Yariv et al., 2020), most approaches require prior knowledge of camera intrinsics from SfM or camera calibration. Our MV-DUSt3R network also processes sparse multi-view input, but does not require prior knowledge of camera parameters.

Neural Scene Reconstruction. Compared to classic methods, which reconstructs a scene using either explicit representations (*e.g.*, 3D point, mesh) or implicit representations (*e.g.*, signed distance function (Newcombe et al., 2011)), recent approaches adopt different neural representations (Sitzmann et al., 2019), including Neural Distance Fields (Chibane et al., 2020; Liu et al., 2020), Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021; Barron et al., 2022; Turki et al., 2022; Kundu et al., 2022; Bian et al., 2023), Gaussian Splatting (Kerbl et al., 2023; Huang et al., 2024; Yu et al., 2024), and their combination (Zou et al., 2024). Many of them require slow per-scene optimization to attain accurate results, while more recent methods explore the use of feed-forward networks for generalizable reconstruction at a fraction of the time, including those for generating Neural Distance Functions (Park et al., 2019; Sitzmann et al., 2020; Chibane et al., 2020), NeRFs (Yu et al., 2021; Du et al., 2023; Chen et al., 2021), and Gaussians (Wewer et al., 2024; Szymanowicz et al., 2024; Charatan et al., 2024; Chen et al., 2024). Note, neural scene reconstruction often requires input views with known camera poses, albeit quite a few exceptions exist, such as CoPoNeRF (Hong et al., 2024), Splatt3R (Smart et al., 2024), and NoPoSplat (Ye et al., 2024). For example, NoPoSplat predicts 3D Gaussians in the same camera coordinates, akin to the key idea of DUSt3R. However, those pose-free methods primarily focus on inference with 2 input views. It is not clear how they perform when processing sparse multi-view input. In contrast, our models MV-DUSt3R, MV-DUSt3R+ equipped with 3D Gaussian splatting heads, not only waive the need for camera pose, but also reconstruct large scenes from multiple views in a single feed-forward pass.

Dense Unconstrained Scene Reconstructions from Multi-View Input. To bypass estimation of camera parameters and poses, recent works like DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) propose a new approach: directly regress pixel-aligned 3D pointmaps for pairs of input views. An expensive 2nd stage global optimization is required to align all pairwise reconstructions in the same coordinate system. Both DUSt3R and MASt3R are only evaluated on object-centric DTU data (Aanaes et al., 2016) where all views are concentrated in a small region. Notably, methods are not validated if their 2-stage pipeline excels at reconstructing larger scenes captured with sparse multi-view input. Subsequently, Spann3R (Wang and Agapito, 2024) augments DUSt3R with a spatial memory to process an ordered set of images. Although capable of performing online scene reconstruction for object-centric scenes, for larger scenes, Spann3R is more likely to drift, generating a misaligned reconstruction due to the limited size of the spatial memory and the lack of globally aligning reconstructions. In contrast, our MV-DUSt3R+ performs offline scene reconstruction by processing all input views (up to 24 in our experiments) at once. Different from DUSt3R, it does not require global optimization because the predicted per-view pointmaps are already globally aligned.

Generative models for 3D reconstruction. Reconstructing scenes from a small number of views is challenging, in particular for unseen areas. Recent advances such as InFusion (Liu et al., 2024b), ZeroNVS (Sargent et al., 2023), Reconfusion (Wu et al., 2024), and ReconX (Liu et al., 2024a) exploit priors encoded in image and video generative models (Blattmann et al., 2023b,a; Song et al., 2023). We leave benefit from image priors of diffusion models as future work.

3 Method

Our goal is to densely reconstruct a scene given a sparse set of rgb images with unknown camera intrinsics and poses. Following DUSt3R, our model predicts 3D pointmaps aligned with 2D pixels for each view. Different from DUSt3R, our model jointly predicts 3D pointmaps for any number of input views in a single forward pass. Formally, given N input image views of a scene $\{I^v\}_{v=1}^N$, where $I^v \in \mathbb{R}^{H \times W \times 3}$, from which we select one reference $r \in \{1, \dots, N\}$, our goal is to predict per-view 3D pointmaps $\{X^{v,r}\}_{v=1}^N$. Note, the 3D pointmap $X^{v,r} \in \mathbb{R}^{H \times W \times 3}$ denotes the coordinates of 3D points for image I^v in the camera coordinate system of the reference view r .

In section 3.1, we introduce our Multi-View Dense Unconstrained Stereo 3D Reconstruction (**MV-DUSt3R**) network to efficiently processes all input views in one pass and without subsequent global optimization, while considering a single chosen reference view. In section 3.2, we present **MV-DUSt3R+**, which processes all input

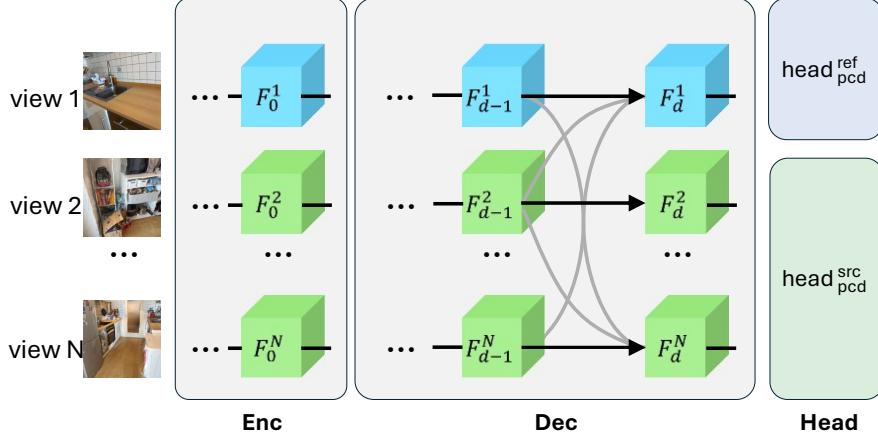


Figure 3 Overview of MV-DUSt3R. Visual tokens for the reference view and other source views are shown in **Blue** and **Green**. Black straight solid lines indicate the primary token flow while gray lines indicate secondary token flow.

views while considering multiple reference views. Finally, to support novel view synthesis, in [section 3.3](#), we augment our networks with Gaussian heads to predict pixel-aligned 3D Gaussians.

3.1 MV-DUSt3R

A Multi-View Model Architecture. As shown in [figure 3](#), MV-DUSt3R consists of an encoder to transform images into visual tokens, decoder blocks to fuse tokens across views, and regression heads to predict per-view 3D pointmaps aligned with 2D pixels. Different from DUSt3R, our network uses decoder blocks to fuse tokens across *all* views rather than independently fusing only tokens for two views at a time. Concretely, a ViT ([Dosovitskiy, 2020](#)) encoder with shared weights, denoted as **Enc**, is first applied on input views $\{I^v\}_{v=1}^N$ to compute initial visual tokens $\{F_0^v\}_{v=1}^N$, *i.e.*, $F_0^v = \text{Enc}(I^v)$. Note, the resolution of the encoder output features is $16\times$ smaller than the input image before being flattened into a sequence of tokens.

To fuse the tokens, two types of decoders are used, one for the chosen reference view and one for the remaining source views. They share the same architecture but their weights differ. Each decoder consists of D decoder blocks referred to as $\text{DecBlock}_d^{\text{ref}}$ and $\text{DecBlock}_d^{\text{src}}$ for $d \in \{1, \dots, D\}$. Their difference is, $\text{DecBlock}_d^{\text{ref}}$ is dedicated to update reference view tokens F^r , while $\text{DecBlock}_d^{\text{src}}$ updates tokens $\{F^v\}_{v \neq r}$ from *all other* source views. Each decoder block takes as input a set of primary tokens from one view, and a set of secondary tokens from other views. In each block, a self-attention layer is applied to primary tokens only, and a cross-attention layer fuses primary tokens with secondary tokens before a final MLP is applied on the primary tokens. Layer norm is also applied before both attentions and the MLP. Using those, the decoder computes the final token representations F_D^v via

$$F_D^v = \begin{cases} \text{DecBlock}_d^{\text{ref}}(F_{d-1}^v, \mathcal{F}_{d-1}^{-v}) & \text{if } v = r, \\ \text{DecBlock}_d^{\text{src}}(F_{d-1}^v, \mathcal{F}_{d-1}^{-v}) & \text{otherwise.} \end{cases} \quad (1)$$

Here, the secondary tokens $\mathcal{F}_d^{-v} = \{F_d^1, \dots, F_d^{v-1}, F_d^{v+1}, \dots, F_d^N\}$ subsume tokens from all views other than the view of the primary tokens F_d^v .

To finally predict the per-view 3D pointmaps, we use two heads: $\text{Head}_{\text{pcd}}^{\text{ref}}$ for the reference view and $\text{Head}_{\text{pcd}}^{\text{src}}$ for all other views. They share the same architecture but use different weights. Each consists of a linear projection layer and a pixel shuffle layer with an upscale factor of 16 to restore the original input image resolution. As in DUSt3R, the head predicts 3D pointmaps $X^{v,r} \in \mathbb{R}^{H \times W \times 3}$ and confidence maps $C^{v,r} \in \mathbb{R}^{H \times W}$ via

$$X^{v,r}, C^{v,r} = \begin{cases} \text{Head}_{\text{pcd}}^{\text{ref}}(F_D^v) & \text{if } v = r, \\ \text{Head}_{\text{pcd}}^{\text{src}}(F_D^v) & \text{otherwise.} \end{cases} \quad (2)$$

Note that DUSt3R is a special case of MV-DUSt3R if the number of views $N = 2$. However, for multiple input views, MV-DUSt3R will update primary tokens using a much larger set of secondary tokens. Hence, it is able to benefit from many more views. Importantly, as our architecture components and structure only

differ slightly from those in DUST3R (additional skip connection and conv net), we have only marginally more trainable parameters. Since, the number of parameters in MV-DUST3R is almost identical to DUST3R, MV-DUST3R can beneficially be initialized using pre-trained DUST3R weights.

Training Recipe. Inspired by DUST3R, we use a confidence-aware pointmap regression loss $\mathcal{L}_{\text{conf}}$, *i.e.*,

$$\mathcal{L}_{\text{conf}} = \sum_{v \in \{1, \dots, N\}} \sum_{p \in P^v} C_p^{v,r} \ell_{\text{regr}}(v, p) - \beta \log C_p^{v,r}, \quad (3a)$$

$$\text{where } \ell_{\text{regr}}(v, p) = \left\| \frac{1}{z} X_p^{v,r} - \frac{1}{\bar{z}} \bar{X}_p^{v,r} \right\|. \quad (3b)$$

Here, P_v denotes the set of valid pixels in view v where groundtruth 3D points are well defined. β controls the weight of the regularization term. The pointmap regression loss ℓ_{regr} measures the difference between predicted and groundtruth 3D points after normalization, which is needed to resolve the scale ambiguity between prediction and groundtruth. It uses $\bar{X}_p^{v,r}$, the groundtruth 3D point of pixel p of view v in the reference view r . The scale normalization factor $z = \text{norm}(\mathcal{X}^{\{v\},r})$ and $\bar{z} = \text{norm}(\bar{\mathcal{X}}^{\{v\},r})$ are computed as the average distance of valid 3D points to the coordinate origin in all views, for prediction and groundtruth, respectively.

3.2 MV-DUST3R+

As shown in [figure 4](#), for different reference view choices, the quality of the scene reconstructed by MV-DUST3R varies spatially. The predicted pointmap for an input source view tends to be better when the viewpoint change to the reference view is small, and deteriorates as the viewpoint change increases. However, to reconstruct a large scene with a sparse set of input views, a single reference view with only moderate viewpoint changes to all other source views is unlikely to exist. Therefore, it is difficult to reconstruct scene geometry equally well everywhere with a single selected reference view. To address this, we propose **MV-DUST3R+**, which selects multiple views as the reference view, and jointly predicts pointmaps for all input views in the camera coordinate of each selected reference view. We hypothesize: while pointmaps of certain input views are difficult to predict for one reference view, they are easier to predict for a different reference view (*e.g.*, smaller viewpoint change, more salient matching patterns). To holistically improve the pointmap prediction of all input views, we include a novel Cross-Reference-View block into MV-DUST3R+.

A Multi-Path Model Architecture. Let $R = \{r^m\}_{m=1}^M$ denote a set of M reference views randomly chosen from an unordered set of input views. We adopt the same decoder blocks from MV-DUST3R, deploy them in a multi-path model architecture (see [figure 5](#)), and use them to compute a reference-view dependent intermediate representation $G_d^{v,m}$ at decoder layer d for input view v and reference view r^m :

$$G_d^{v,m} = \begin{cases} \text{DecBlock}_d^{\text{ref}}(F_{d-1}^{v,m}, \mathcal{F}_{d-1}^{-v,m}) & \text{if } v = r^m, \\ \text{DecBlock}_d^{\text{src}}(F_{d-1}^{v,m}, \mathcal{F}_{d-1}^{-v,m}) & \text{otherwise,} \end{cases} \quad (4)$$

$$F_d^{v,m} = \text{CrossRefViewBlock}_d(G_d^{v,m}, \mathcal{G}_d^{v,-m}). \quad (6)$$

Here, $\mathcal{F}_{d-1}^{-v,m} = \{F_{d-1}^{1,m}, \dots, F_{d-1}^{v-1,m}, F_{d-1}^{v+1,m}, \dots, F_{d-1}^{N,m}\}$. As shown in [figure 5](#), we fuse and update per-view tokens computed under different reference views by adding a *Cross-Reference-View* block after each decoder block ([equation \(6\)](#)), where $\mathcal{G}_d^{v,-m} = \{G_d^{v,1}, \dots, G_d^{v,m-1}, G_d^{v,m+1}, \dots, G_d^{v,M}\}$. Following [equation \(2\)](#), we compute per-view pointmaps $X^{v,m}$ and confidence map $C^{v,m}$ under each reference view r^m .

Training Recipe. Compared with MV-DUST3R, MV-DUST3R+ only adds a small number of additional trainable parameters via the Cross-Reference-View blocks (see appendix). During training, a random subset of M input views are selected as the reference views. We average the pointmap regression losses in [equation \(3a\)](#) for all reference views.

Model Inference. At inference time, we uniformly select a subset of M input views as the reference views, while the 1st input view is always selected. A model with M paths is used but the final per-view pointmap predictions are computed using the heads in the 1st path.

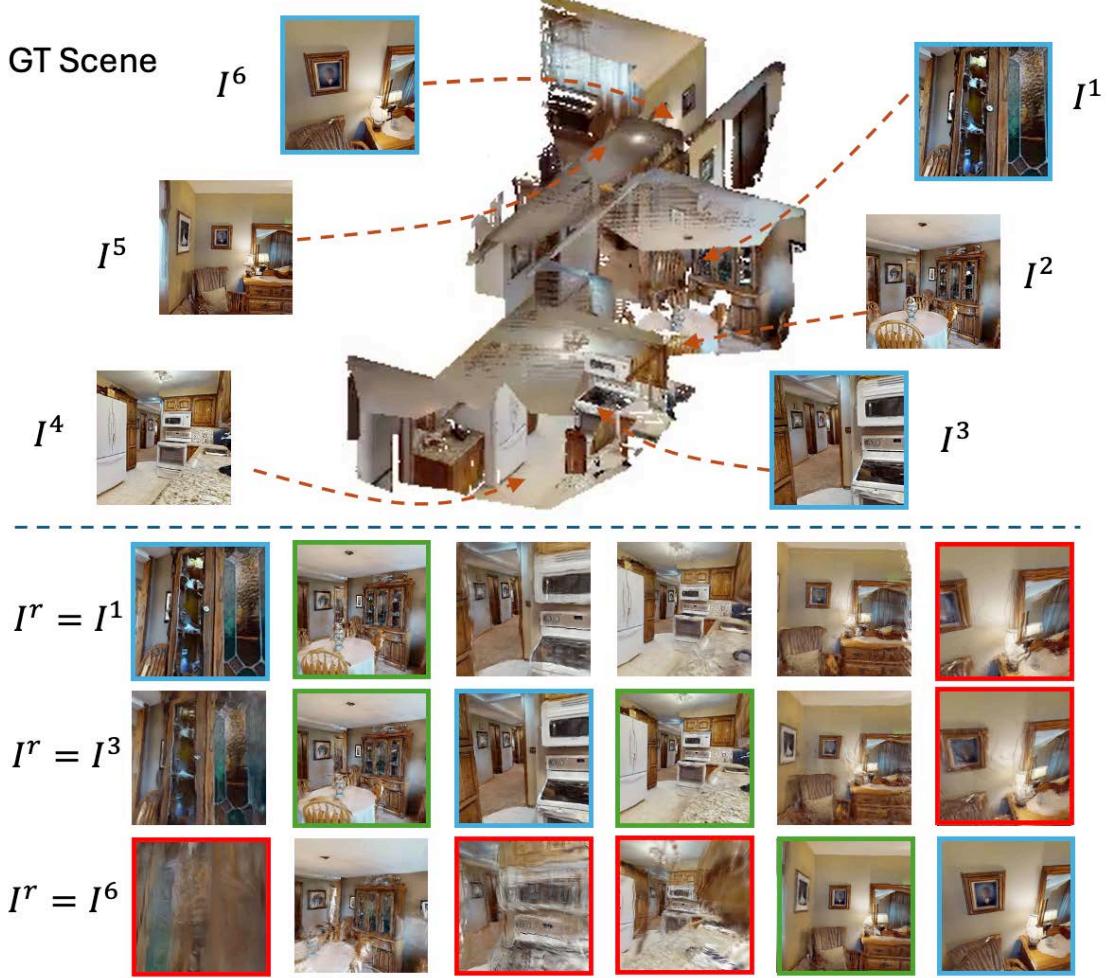


Figure 4 Top: A multi-room scene: 16 views are sampled as input to MV-DUSt3R. For clarity, only 6 are shown. 3 of them are reference view candidates, highlighted in blue. **Bottom:** In each row, we select a different reference view and render the reconstructed scene from 6 input views. Renderings in good and poor quality are highlighted in green and red. As the viewpoint change between the input view and the reference view increases, quality of the reconstructed scene geometry in that input view decreases.

3.3 MV-DUSt3R(+) for Novel View Synthesis

Next we extend our networks to support NVS with Gaussian primitives (Kerbl et al., 2023). For clarity, below we use MV-DUSt3R+ as an example. MV-DUSt3R can be extended similarly.

Gaussian Head. We add a separate set of heads to predict per-pixel Gaussian parameters, including scaling factor $S^{v,m} \in \mathbb{R}^{H \times W \times 3}$, rotation quaternion $q^{v,m} \in \mathbb{R}^{H \times W \times 4}$, and opacity $\alpha^{v,m} \in \mathbb{R}^{H \times W}$. We add Gaussian heads $\text{Head}_{3DGS}^{\text{ref}}$ and $\text{Head}_{3DGS}^{\text{src}}$ for reference and other views:

$$S^{v,m}, q^{v,m}, \alpha^{v,m} = \begin{cases} \text{Head}_{3DGS}^{\text{ref}}(F_D^{v,m}) & \text{if } v = r^m, \\ \text{Head}_{3DGS}^{\text{src}}(F_D^{v,m}) & \text{otherwise.} \end{cases} \quad (7)$$

$$(8)$$

For other Gaussian parameters, we use the predicted pointmap $X^{v,m}$ as the center, the pixel color I^v as the color and fix the spherical harmonics degree to be 0.

Training Recipe. During training, for a chosen reference view r^m , we perform differentiable splatting-based rendering (Kerbl et al., 2023) to generate rendering predictions for both input views and novel views. Following prior approaches (Szymanowicz et al., 2024; Charatan et al., 2024; Chen et al., 2024), we use a weighted

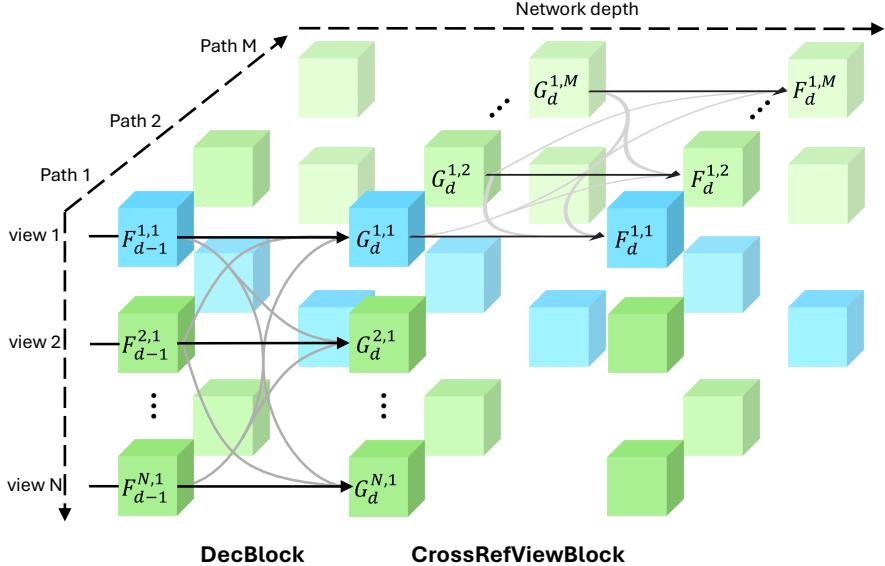


Figure 5 DecBlock and CrossRefViewBlock in MV-DUST3R+: tokens of the reference and other views are highlighted in blue and green, respectively. Each model path uses a different reference view. For clarity, only 1 of stacked DecBlock and CrossRefViewBlock are shown.

Dataset	Eval setting	Scene type
HM3D	Supervised	multi-room (large)
ScanNet	Supervised	single-room (small)
MP3D	Zero-shot	multi-room & outdoor (largest)

Table 1 Evaluation datasets comparisons.

sum of L^2 pixel difference loss and perceptual similarity loss LPIPS as the rendering loss $\mathcal{L}_{\text{render}}$ to train the Gaussian heads. The final training loss includes both $\mathcal{L}_{\text{conf}}$ and $\mathcal{L}_{\text{render}}$ (for details see appendix).

4 Experiments

4.1 Datasets

Our training data includes ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), HM3D (Ramakrishnan et al., 2021), and Gibson (Xia et al., 2018). Note, all of them are also used by DUST3R. For evaluation, we use datasets MP3D (Chang et al., 2017), HM3D (Ramakrishnan et al., 2021), and ScanNet (Dai et al., 2017). While ScanNet scenes are often small single-room sized and with low diversity, scenes in MP3D and HM3D are often large multi-room sized and with high diversity. MP3D also contains outdoor scenes. See table 1 to compare evaluation datasets. We use the same train/test split as DUST3R, and our training data is a subset of DUST3R’s training data (for details see appendix).

Trajectory Generation. To generate a set of input views $\{I^v\}_{v=1}^N$ for $N > 2$, we first randomly select one frame and initialize the current scene point cloud using its data. Then we sequentially sample more candidate frames. We retain a candidate frame and add its corresponding point cloud to the current scene, if the overlap between the candidate frame’s point cloud and the current scene point cloud is between a lower threshold t_{\min} and an upper bound t_{\max} .

Training Trajectories. To sample the training set trajectories, we employ two choices of thresholds: $(t_{\min}, t_{\max}) \in \{(30\%, 70\%), (30\%, 100\%)\}$. From ScanNet and ScanNet++, we sample 1K trajectories of 10 views per scene, and a total of 3.2M trajectories. On HM3D and Gibson, where the scene is often larger, we sample 6K trajectories per scene with 10 views each, and a total of 7.8M trajectories.

	Method	GO	HM3D			ScanNet			MP3D			Time (sec)
			ND ↓	DAc ↑	CD ↓	ND ↓	DAc ↑	CD ↓	ND ↓	DAc ↑	CD ↓	
4 views	Spann3R	✗	37.1	0.0	225(184)	8.9	19.5	54.7(50.1)	42.7	0.0	248(202)	0.36
	DUSt3R	✓	1.9	75.1	5.6(2.3)	1.3	89.8	4.0(0.4)	3.9	41.7	40.0(5.3)	2.42
	MV-DUSt3R	✗	1.1	92.2	2.0(1.1)	1.0	93.3	2.0(0.4)	2.5	62.4	25.3(4.1)	0.05
	MV-DUSt3R+	✗	1.0	95.2	1.5(0.9)	0.8	94.9	1.5(0.3)	2.2	68.0	19.9(3.4)	0.29
	MV-DUSt3R _{oracle}	✗	1.0	94.6	1.5(0.7)	0.8	95.5	1.3(0.3)	2.3	66.6	20.7(4.0)	-
	MV-DUSt3R+ _{oracle}	✗	0.9	96.5	1.4(0.7)	0.7	95.8	1.2(0.2)	2.1	70.6	17.9(3.3)	-
12 views	Spann3R	✗	32.6	0.0	125(113)	9.1	16.3	36.6(31.2)	35.0	0.0	138(112)	1.34
	DUSt3R	✓	3.9	30.7	18.1(3.4)	1.9	82.6	4.1(0.6)	6.6	12.0	49.6(8.3)	8.28
	MV-DUSt3R	✗	1.6	79.5	3.0(1.2)	1.4	86.8	2.3(0.8)	3.4	41.3	22.6(5.5)	0.15
	MV-DUSt3R+	✗	1.2	91.5	1.8(0.7)	1.2	88.4	1.8(0.7)	2.6	55.0	15.1(3.8)	0.89
	MV-DUSt3R _{oracle}	✗	1.3	88.8	1.8(0.9)	1.0	90.6	1.3(0.7)	2.9	51.3	16.4(4.0)	-
	MV-DUSt3R+ _{oracle}	✗	1.1	94.8	1.4(0.7)	1.0	90.9	1.3(0.5)	2.5	59.8	13.6(3.5)	-
24 views	Spann3R	✗	41.7	0.0	139(121)	11.4	1.6	37.4(35.5)	46.6	0.0	151(121)	2.73
	DUSt3R	✓	6.8	7.3	32.4(5.2)	2.4	72.6	5.1(1.0)	11.4	2.5	80.9(14.3)	27.21
	MV-DUSt3R	✗	3.4	36.7	10.0(3.5)	2.2	75.2	2.7(0.9)	6.3	12.2	38.6(13.9)	0.35
	MV-DUSt3R+	✗	2.1	64.5	3.9(2.0)	1.6	81.2	1.7(0.7)	4.3	26.7	22.0(5.9)	1.97
	MV-DUSt3R _{oracle}	✗	2.1	58.9	3.5(2.1)	1.4	82.9	1.4(0.7)	4.4	22.0	19.9(5.1)	-
	MV-DUSt3R+ _{oracle}	✗	1.8	77.9	2.6(1.3)	1.3	85.1	1.3(0.6)	3.6	33.1	15.1(4.4)	-

Table 2 MVS reconstruction results. Best results are marked in red. For clarity, metric ND is scaled up by 10×, DAc is in percent %, and CD is scaled by 100× with its median given in parentheses. Results of our **oracle** methods are obtained by manually selecting the best single reference view from reference view candidates to report a possibly achievable performance. In the rightmost column, each method’s time for scene reconstruction is reported.

Test Trajectories. For the test set, we generate 1K trajectories per dataset. To support evaluation with a larger number of inputs views, we sample 30 views per trajectory.

4.2 Implementation Details

We process input views at resolution 224×224 . We utilize 64 Nvidia H100 GPUs for the model training. To initialize, DUSt3R model weights are used. We use the first $N = 8$ views of each trajectory as input views, and randomly select 1 view as the reference view for MV-DUSt3R and $M = 4$ views for MV-DUSt3R+. We train for 100 epochs using 150K trajectories per epoch, which takes 180 hours. For MVS reconstruction evaluation, to assess the performance of each method in reconstructing scenes of variable sizes, we report results with input views ranging from 4 to 24 views. For NVS evaluation, we use the remaining 6 views as novel views. Below we report results on all evaluation datasets for all choices of the number of input views using only one MV-DUSt3R model and one MV-DUSt3R+ model.

4.3 Multi-View Stereo Reconstruction

Metrics. We report Chamfer Distance (CD), as in prior work (Aanæs et al., 2016; Wang et al., 2024), as well as 2 additional metrics. **NormalizedDistance (ND):** ℓ_{regr} with zero-centering to make it scale and translation invariant; **DistanceAccu@0.2 (DAc):** proportion of pixels where the corresponding normalized distance between prediction and groundtruth in pointmap is ≤ 0.2 .

Baselines. We compare with baselines that reconstruct scenes from input rgb views without knowing camera intrinsics and poses. We evaluate the DUSt3R model trained at input resolution 224×224 with Global Optimization (GO), and also the Spann3R (Wang and Agapito, 2024) model on Github.

Results are shown in table 2. In the supervised setting, we compare DUSt3R and our methods on HM3D and ScanNet. On HM3D (multi-room scenes), **MV-DUSt3R** consistently outperforms DUSt3R, as the scene size increases and more input views are sampled (from 4 to 24 views). For example, MV-DUSt3R reduces ND by 1.7× and increases DAc by 1.2× for 4-view input. For 24-view input, MV-DUSt3R improves ND by 2× and DAc by 5.3×. This confirms: as more input views are available, single-stage MV-DUSt3R exploits multi-view cues to infer 3D scene geometry better than DUSt3R, which only exploits pairwise stereo cues

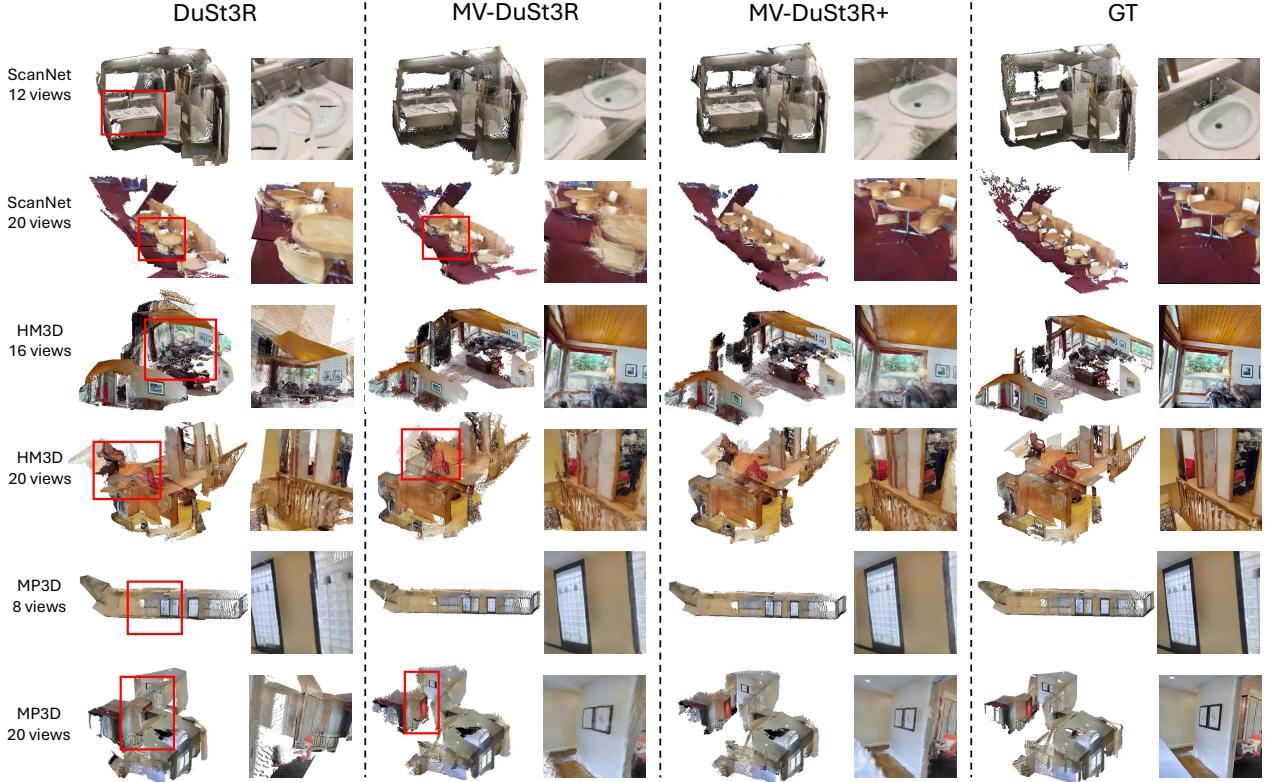


Figure 6 MVS reconstruction and NVS qualitative results. We show one method in each column, which includes the reconstructed pointcloud and 1 rendered new view. Incorrectly reconstructed geometry is highlighted in red boxes. DUSt3R often introduces incorrect pairwise reconstructions when the scene has multiple objects with similar appearance (e.g., windows, chairs, doors), which can not be recovered by the global optimization. MV-DUSt3R is more robust overall but still sometimes fails to reconstruct geometry accurately in regions far away from the reference view, while MV-DUSt3R+ predicts geometry more evenly across the space.

at a time. Furthermore, **MV-DUSt3R+** substantially improves upon MV-DUSt3R, especially when the scene size is large and many more input views are used. With 12-view input, MV-DUSt3R+ improves ND by $1.3\times$ and DAc by $1.2\times$. With 24-view input, the improvements are more significant, with a $1.6\times$ lower ND and a $1.8\times$ higher DAc. The multi-path architecture enables MV-DUSt3R+ to more effectively fuse multi-view cues across different choices of reference frames, which holistically improves the scene geometry reconstruction in all input views. For a qualitative comparisons, see figure 6. For zero-shot evaluation on the challenging MP3D data, all methods have worse results. However, both of our methods consistently improve upon DUSt3R across different numbers of input views.

In all settings, Spann3R performs worse than all other methods, and often fails to reconstruct a scene from sparse views, a setting which is more challenging than the sequential video frames used for evaluation in the original paper.

4.4 Multi-View Pose Estimation

For both baseline and our methods, we estimate the relative camera pose for all pairs of input views from a given set of input views. We use the Weiszfeld algorithm (Plastria, 2011) to estimate camera intrinsics, and RANSAC (Fischler and Bolles, 1981) with PnP (Lepetit et al., 2009) to estimate camera pose (see appendix for more details).

Baselines. We compare with other pose-free methods including DUSt3R and PoseDiffusion (Wang et al., 2023b), a recent diffusion based method for camera pose estimation.

Metrics. Prior methods (Wang et al., 2024) report Relative Rotation Accuracy (**RRA@15**), Relative Translation

	Method	GO	HM3D			ScanNet			MP3D		
			RRE ↓	RTE ↓	mAE ↓	RRE ↓	RTE ↓	mAE ↓	RRE ↓	RTE ↓	mAE ↓
4 views	DUSt3R	✓	2.4	3.1	12.5	3.0	20.0	30.7	3.5	3.8	13.3
	MV-DUSt3R	✗	1.5	1.5	5.5	2.3	16.8	27.0	1.2	1.0	5.4
	MV-DUSt3R+	✗	1.2	1.1	4.9	1.4	16.1	26.2	0.8	0.8	4.6
	MV-DUSt3R _{oracle}	✗	0.0	0.1	2.8	0.9	7.0	18.9	0.1	0.1	2.9
	MV-DUSt3R+ _{oracle}	✗	0.0	0.0	2.4	0.9	6.8	18.7	0.1	0.0	2.4
12 views	DUSt3R	✓	3.7	8.3	20.1	4.6	22.6	34.2	4.5	8.4	19.8
	MV-DUSt3R	✗	1.5	2.6	8.4	3.7	14.7	26.1	1.6	2.6	8.2
	MV-DUSt3R+	✗	0.6	1.2	5.2	2.5	11.6	22.9	0.5	1.0	4.9
	MV-DUSt3R _{oracle}	✗	0.4	0.6	4.9	1.7	7.8	20.2	0.6	0.7	5.1
	MV-DUSt3R+ _{oracle}	✗	0.3	0.3	3.4	1.7	6.0	17.9	0.3	0.3	3.3
24 views	DUSt3R	✓	8.8	18.1	30.9	8.1	26.6	38.9	10.0	18.2	30.5
	MV-DUSt3R	✗	8.9	12.8	23.7	8.2	21.9	34.2	8.2	11.1	21.4
	MV-DUSt3R+	✗	3.0	6.5	15.8	4.6	16.7	29.4	3.3	6.0	14.6
	MV-DUSt3R _{oracle}	✗	3.2	4.4	14.7	3.4	13.1	26.7	3.4	4.2	14.0
	MV-DUSt3R+ _{oracle}	✗	1.4	2.4	11.1	2.6	9.9	23.7	1.8	2.4	10.6

Table 3 Multi-View Pose Estimation results. Metrics are reported in percent %.

Accuracy (**RTA@15**) under threshold 15 degrees, and mean Average Accuracy (**mAA@30**) under threshold 30 degrees. For clarity, we report Relative Rotation Error (**RRE@15** = 1.0 – RRA@15), Relative Translation Error (**RTE@15** = 1.0 – RTA@15) and mean Average Error (**mAE@30** = 1.0 – mAA@30).

Results. The comparisons with DUSt3R are presented in [table 3](#), while comparisons with PoseDiffusion are included in the appendix, as we find PoseDiffusion significantly underperforms other methods on our evaluation datasets. As shown in [table 3](#), in the supervised setting on HM3D, our method MV-DUSt3R achieves a 2.3× lower mAE for 4-view input, and a 1.3× lower mAE for 24-view input, compared with DUSt3R. MV-DUSt3R+ performs best, achieving a 2.6× lower mAE for 4-view input, and a 2.0× lower mAE for 24-view input, compared with DUSt3R. For another supervised setting ScanNet and the zero-shot MP3D, our method MV-DUSt3R+ performs best, while MV-DUSt3R consistently outperforms DUSt3R under all circumstances.

4.5 Novel View Synthesis

Metrics. Following prior works ([Smart et al., 2024](#); [Charatan et al., 2024](#); [Fan et al., 2024](#)), we report Peak Signal-to-Noise Ratio (**PSNR**), Structural Similarity Index Measure (**SSIM**) ([Wang et al., 2004](#)), and Learned Perceptual Image Patch Similarity (**LPIPS**) ([Zhang et al., 2018](#)).

Baseline. We compare with a DUSt3R-based baseline, which generates per-pixel Gaussian parameters as follows. We use the pointmap predicted by DUSt3R as the Gaussian center, use pixel RGB color I^v as the color, a constant 0.001 for the scale factor $S^{v,m}$, an identity transform, 1.0 for opacity, and spherical harmonics with zero-degree. See appendix for more details on rendering.

Results. As shown in [table 4](#), MV-DUSt3R improves upon the DUSt3R baseline across all evaluation datasets under all choices of input views. The improvements are also confirmed qualitatively in [figure 6](#): the novel views synthesized by MV-DUSt3R better infer 3D geometry of objects and background (*e.g.*, walls, ceiling). MV-DUSt3R+ further improves in challenging situations, such as a scene with multiple close-by objects of similar appearance (*e.g.*, chairs). As an example, consider the ScanNet scene with 20 views in [figure 6](#). Using multiple reference views, and fusing features computed in different model paths help resolve ambiguity in inferring the spatial relations between input views.

4.6 Scene Reconstruction Time

We compare the time of MVS reconstruction in [table 2](#). Our single-stage feed-forward networks entirely run on a GPU, without Global Optimization (GO). Compared with DUSt3R, our MV-DUSt3R runs 48× to 78×

	Method	GO	HM3D			ScanNet			MP3D		
			PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
4 views	DUSt3R	✓	16.0	5.0	3.7	17.0	6.0	3.0	15.5	4.6	4.0
	MV-DUSt3R	✗	19.9	6.0	2.0	21.9	7.1	1.6	19.6	5.8	2.1
	MV-DUSt3R+	✗	20.2	6.1	1.9	22.2	7.1	1.5	19.9	5.9	2.0
	MV-DUSt3R _{oracle}	✗	21.0	6.5	1.6	22.8	7.4	1.4	20.6	6.2	1.8
	MV-DUSt3R+ _{oracle}	✗	21.4	6.6	1.5	23.0	7.4	1.4	21.0	6.3	1.7
12 views	DUSt3R	✓	15.1	4.4	4.9	16.3	5.4	3.6	14.7	4.0	5.3
	MV-DUSt3R	✗	18.9	5.6	2.7	20.1	6.5	2.2	18.4	5.3	2.8
	MV-DUSt3R+	✗	19.4	5.8	2.4	20.4	6.6	2.1	19.0	5.5	2.6
	MV-DUSt3R _{oracle}	✗	19.9	5.9	2.2	21.0	6.8	1.9	19.3	5.6	2.4
	MV-DUSt3R+ _{oracle}	✗	20.4	6.1	2.0	21.3	6.8	1.8	19.9	5.8	2.2
24 views	DUSt3R	✓	14.3	4.2	5.6	15.2	5.0	4.2	13.8	3.6	6.1
	MV-DUSt3R	✗	17.8	5.3	3.6	18.4	6.0	2.9	17.3	4.9	3.8
	MV-DUSt3R+	✗	18.4	5.4	3.2	18.6	6.0	2.8	17.9	5.1	3.5
	MV-DUSt3R _{oracle}	✗	18.5	5.5	3.1	19.3	6.2	2.5	18.0	5.1	3.3
	MV-DUSt3R+ _{oracle}	✗	19.0	5.6	2.9	19.4	6.2	2.5	18.5	5.3	3.1

Table 4 Novel View Synthesis results. For clarity, we scale up metrics SSIM and LPIPS by 10 \times .

Test	Training recipe	MVS Reconstruction			MVPE			NVS		
		ND \downarrow	DAc \uparrow	CD \downarrow	RRE \downarrow	RTE \downarrow	mAE \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
4 views	1-stage, 4 views	1.0	94.4	1.7(1.2)	0.9	0.9	2.6	20.7	6.3	1.7
	1-stage, 8 views	1.0	95.2	1.5(0.9)	1.2	1.1	4.9	20.2	6.1	1.9
	2-stage, mixed views	0.9	95.5	1.5(0.8)	0.8	0.7	2.0	20.7	6.3	1.8
12 views	1-stage, 4 views	6.3	0.3	23.8(18.2)	15.1	17.5	34.1	16.5	4.9	4.4
	1-stage, 8 views	1.2	91.5	1.8(0.7)	0.6	1.2	5.2	19.4	5.8	2.4
	2-stage, mixed views	1.2	92.2	1.5(1.0)	0.4	0.8	3.8	19.5	5.9	2.2
24 views	1-stage, 4 views	17.7	0.0	81.4(55.5)	45.5	47.4	63.2	14.5	4.6	6.2
	1-stage, 8 views	2.1	64.5	3.9(2.0)	3.0	6.5	15.8	18.4	5.4	3.2
	2-stage, mixed views	1.7	81.4	2.6(1.3)	1.4	3.0	9.1	19.1	5.7	2.7

Table 5 Impact of # of input views at training time on the performance of MV-DUSt3R+ on the HM3D evaluation set.

faster than DUSt3R, while the more performant MV-DUSt3R+ runs 8 \times to 14 \times faster, when considering 4 to 24 input views. *MV-DUSt3R+ reconstructs 24-view input for scenes of average size 17.9 m² on HM3D and 37.3 m² on MP3D in less than 2 seconds.*

4.7 Ablation Studies

Number of input views at training time. We compare 1-stage and 2-stage trained MV-DUSt3R+ on the HM3D evaluation set. For 1-stage training, we choose the first 4 or 8 views of the trajectory. For 2-stage training, we finetune upon MV-DUSt3R+'s 1-stage 8-view training, by using a mixed set of inputs with the number of views uniformly sampled between 4 and 12. As shown in [table 5](#), 1-stage training on 4 views doesn't generalize well to more views, 1-stage training on 8 views performs decent, and 2-stage training outperforms 1-stage training on almost all tasks on HM3D. See appendix for more 2-stage results.

Upper bound performance of our networks. We study oracle performance if the best reference view is chosen based on groundtruth. For MV-DUSt3R, we consider all input views as reference view candidates, and manually select the one with best MVS reconstruction (**MV-DUSt3R_{oracle}**). For MV-DUSt3R+, we choose the model path with best MVS reconstruction (**MV-DUSt3R+_{oracle}**). As shown in [tables 2 to 4](#), the performance gap between MV-DUSt3R+ and MV-DUSt3R+_{oracle} is significantly smaller than that between MV-DUSt3R and MV-DUSt3R_{oracle}. This validates our multi-path MV-DUSt3R+ architecture.

Impact of adding Gaussian head on MVS reconstruction performance. In [table 6](#), we compare MV-DUSt3R and

Method	GS	HM3D			ScanNet			MP3D		
		ND ↓	DAc ↑	CD ↓	ND ↓	DAc ↑	CD ↓	ND ↓	DAc ↑	CD ↓
MV-DUSt3R	✗	1.1	93.5	1.9(1.4)	1.0	93.3	2.0(0.4)	2.6	61.6	25.4(4.9)
	✓	1.1	92.2	2.0(1.1)	1.0	93.3	2.0(0.4)	2.5	62.4	25.3(4.1)
MV-DUSt3R+	✗	0.9	95.2	1.5(1.1)	0.8	94.8	1.4(0.4)	2.2	68.5	18.9(3.5)
	✓	1.0	95.2	1.5(0.9)	0.8	94.9	1.5(0.3)	2.2	68.0	19.9(3.4)
MV-DUSt3R	✗	3.3	37.6	10.0(5.1)	2.2	75.8	2.7(0.7)	6.4	11.0	43.3(13.0)
	✓	3.4	36.7	10.0(3.5)	2.2	75.2	2.7(0.9)	6.3	12.2	38.6(13.9)
MV-DUSt3R+	✗	2.1	68.1	4.4(2.5)	1.4	84.9	1.5(0.5)	4.3	26.5	22.6(5.6)
	✓	2.1	64.5	3.9(2.0)	1.6	81.2	1.7(0.7)	4.3	26.7	22.0(5.9)

Table 6 Impact of adding Gaussian (GS) heads on MVS reconstruction performance on HM3D, ScanNet, and MP3D datasets.

MV-DUSt3R+ models with and without Gaussian heads, while keeping other settings identical. As shown in [table 6](#), adding Gaussian heads does not significantly improve or degrade performance of our models on MVS reconstruction.

5 Conclusion

We propose fast single-stage networks MV-DUSt3R and MV-DUSt3R+ to reconstruct scenes from up to 24 input views in one feed-forward pass without requiring camera intrinsics and poses. We extensively evaluate results on 3 datasets in both supervised and zero-shot settings, and confirm compelling results and efficiency over prior art.

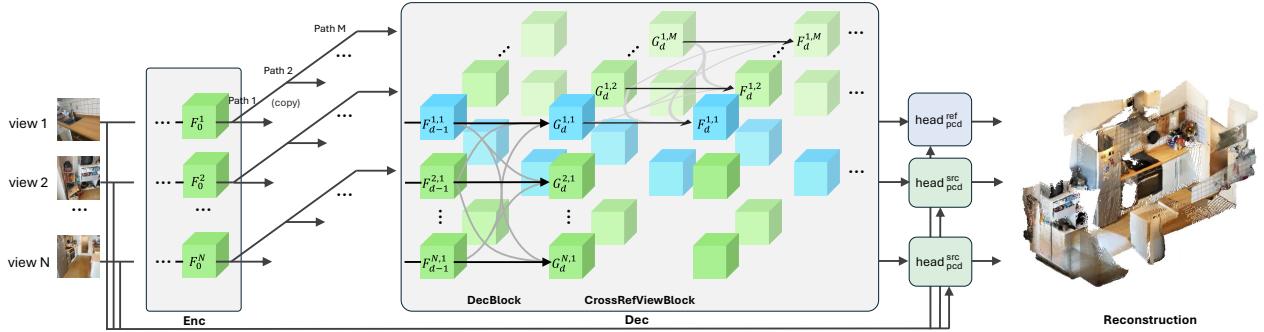


Figure S1 Full pipeline of MV-DUSt3R+ without Gaussian heads. Extra skip connections are sent to the regression head for fine-grained correction.

Appendix

A MV-DUSt3R (+) for Novel View Synthesis

Below we present more implementation details in order to extend MV-DUSt3R+ to the NVS task. Note, MV-DUSt3R can be viewed as a special case of MV-DUSt3R+ with the number of reference view candidates $M = 1$.

Training Recipe. During training, for a chosen reference view $r^m \in R$, we perform differentiable splatting-based rendering (Kerbl et al., 2023) to generate rendering predictions for a set of target views $T = \{t^k\}_{k=1}^{N+N'}$, including both N input views and N' novel views. We transform the Gaussian parameters predicted for input view v in the coordinate system of the reference view r^m into the coordinate system of a target view t^k . For this we use the groundtruth camera poses P^m of the reference view and P^k of the target view. Formally, the Gaussian centers $X^{v,m}$ are transformed as follows:

$$\hat{X}^{v,m \rightarrow k} = P^k(P^m)^{-1} \left(\frac{\bar{z}}{z} X^{v,m} \right). \quad (\text{S1})$$

Note, we account for scale ambiguity between predicted and groundtruth scene, as described in Section 3.2 of the main paper, by scaling the Gaussian center by \bar{z}/z . We use $\hat{X}^{v,m \rightarrow k}$ to denote the transformed coordinates. Other Gaussian parameters, including scaling factor $S^{v,m}$ and rotation quaternions $q^{v,m}$, are scaled or transformed accordingly into $\hat{S}^{v,m}$ and $q^{v,m \rightarrow k}$. For each target view t^k and each reference view candidate r^m , a splatting-based rendering $\hat{I}^{k,m}$ is generated using all N per-input-view pointmaps predicted from the path m in the multi-path model. Formally,

$$\hat{I}^{k,m} = \text{Rendering}(Q^{1,m \rightarrow k}, \dots, Q^{N,m \rightarrow k}), \quad (\text{S2})$$

where $Q^{v,m \rightarrow k} = \{I^v, \hat{X}^{v,m \rightarrow k}, \hat{S}^{v,m}, q^{v,m \rightarrow k}, \alpha^{v,m}\}$ denotes the transformed pixel-aligned Gaussian parameters of input view v .

Following prior approaches (Szymanowicz et al., 2024; Charatan et al., 2024; Chen et al., 2024), we use a weighted sum of an ℓ_2 pixel difference loss and the perceptual similarity loss LPIPS (Zhang et al., 2018) as the rendering loss $\mathcal{L}_{\text{render}}$ to train the Gaussian heads. Formally,

$$\mathcal{L}_{\text{render}} = \frac{1}{|T||M|} \sum_{(k,m)} \left(\left\| \hat{I}^{k,m} - I^k \right\|_2^2 + \gamma \text{LPIPS}(\hat{I}^{k,m}, I^k) \right), \quad (\text{S3})$$

where γ controls the weight of the LPIPS loss. The final loss to train our models with Gaussian heads includes both the confidence-aware pointmap regression loss and the rendering loss, i.e.,

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{conf}} + \delta \mathcal{L}_{\text{render}}. \quad (\text{S4})$$

Here, δ is a weight to balance the two losses.

Module	DUS3R	MV-DUS3R+
Enc	Conv2d(in=3, out=1024, kernel=16, stride=16) EncBlock(embed_dim=1024, n_head=16) * 16	Conv2d(in=3, out=1024, kernel=16, stride=16) EncBlock(embed_dim=1024, n_head=16) * 16
Dec	Linear(in=1024, out=768) DecBlock ^{ref/src} (embed_dim=768, n_head=12) * 12 -	Linear(in=1024, out=768) DecBlock ^{ref/src} (embed_dim=768, n_head=12) * 12 CrossRefViewBlock(embed_dim=768, n_head=12) * 12
Head _{pcd}	Linear(in=768, out=768) PixelShuffle(patch_size=16) - - - -	Linear(in=768, out=768) PixelShuffle(patch_size=16) Conv2d(in=6, out=256, kernel=3, stride=1) Conv2d(in=256, out=256, kernel=5, stride=1) Conv2d(in=256, out=256, kernel=5, stride=1) Conv2d(in=256, out=3, kernel=3, stride=1)
Head _{3DGS}	- -	Linear(in=768, out=768) PixelShuffle(patch_size=16)

Table S1 Architecture Comparisons. New parameters in MV-DUS3R+ are highlighted in red color.

Model Inference during evaluation. The inference of MV-DUS3R+ with Gaussian heads largely follows the MV-DUS3R+ model inference. A subset of M input views are selected as reference views, and the per-view pixel-aligned 3D Gaussian predictions are computed using the heads in the 1st path of the multi-path model. To generate a rendering at a novel view during evaluation, we transform the predicted Gaussian parameters from the 1st reference view to the novel view based on their groundtruth poses, and run splatting-based rendering.

B Implementation Details

B.1 Network Architecture

An overview of the MV-DUS3R+ model is provided in figure S1. Inspired by DUS3R, our new CrossRefViewBlock shares the same block architecture as the DecBlock, which includes a cross-attention block, a self-attention block and a MLP. Its parameters are randomly initialized except that zero-initialization is used for the last layer in the cross attention block, the self attention block and the MLP. Our Gaussian prediction heads share the same structure as DUS3R’s pointmap prediction head. However, our pointmap prediction head differs and has slightly more trainable parameters to improve accuracy.

For clarity, we slightly abuse the notation and remove the superscript of the input view and the reference view below to present the head Head_{pcd}. We start with the linear head design in DUS3R, which consists of a Linear layer followed by a PixelShuffle layer to restore the original resolution. Differently, we improve the vanilla head by adding a skip connection from the input view to the head output to restore more high-resolution details in the final pointmap prediction. The skip connection is implemented as a small ConvNet, which consists of multiple stride-1 convolutional layers:

$$X_c^v = \text{PixelShuffle}(\text{Linear}(F_D^v)), \quad (\text{S5})$$

$$X^v = \text{ConvNet}(X_c^v, I^v) + X_c^v. \quad (\text{S6})$$

Here, X_c^v is the coarse point map and the ConvNet includes 4 sequential stride-1 convolutional layers with kernel size 3, 5, 5, and 3 respectively. See table S1 for more details and a comparison of network architectures and hyperparameters.

B.2 Model Training Recipe

We implement our approach in PyTorch (Paszke et al., 2019). We train our models for 100 epochs using 150K trajectories per epoch, a learning rate of 1.5e-4 with a cosine schedule, and Adam optimizer (Diederik, 2014). Training takes 180 hours. The hyperparameters in the loss functions are $\beta = 0.2$, $\gamma = 1$, and $\delta = 1$. The input image resolution is 224×224 . We use the open-sourced weights of DUS3R¹, trained at the input resolution

¹<https://github.com/naver/dust3r?tab=readme-ov-file#checkpoints>

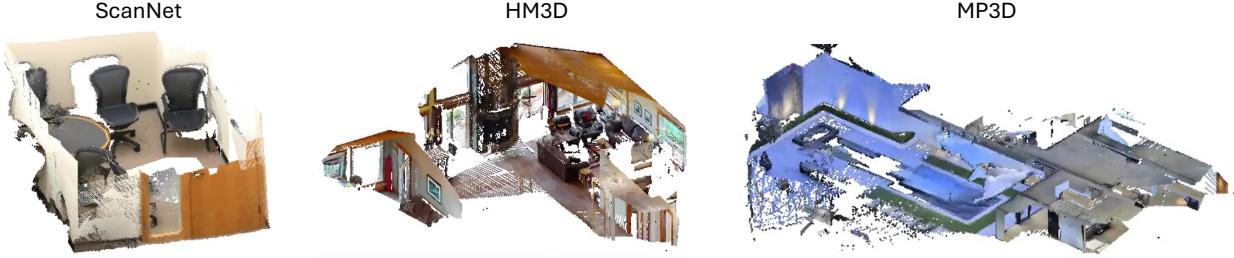


Figure S2 Scenes with 20 sampled input views from the evaluation datasets. ScanNet scenes are small and only contains a single room. HM3D scenes often cover a larger area and multiple rooms. MP3D scenes are on average the largest with the highest diversity, which often not only cover some indoor region, but also the outdoor pool and garden.

224×224 , to initialize all parameters existing within our models. We do not finetune the DUST3R model trained at the higher 512 resolution for two reasons: 1) the lower resolution model is representative enough given its performance on Multi-View depth evaluation is similar to that of the higher resolution model as reported in the original DUST3R paper; and 2) finetuning at the higher 512 resolution is computationally costly. We leave scaling of MV-DUST3R+ training at the higher 512 resolution to future work.

We use the first $N = 8$ views of each 10-view training trajectory as input views. Among those $N = 8$ input views, we select *one* random view as the reference view when training MV-DUST3R and $M = 4$ random reference views when training MV-DUST3R+. The remaining $N' = 2$ views are considered as novel views.

C Experiments

C.1 Datasets

In our experiments, we use ScanNet (Dai et al., 2017) (1221 scenes for training, 292 scenes for test), ScanNet++ (Yeshwanth et al., 2023) (360 scenes for training), HM3D (Ramakrishnan et al., 2021) (800 scenes for training, 100 scenes for test), Gibson (Xia et al., 2018) (500 scenes for training) and MP3D (Chang et al., 2017) (no scenes for training, 18 scenes for test). Note that our 3 evaluation datasets differ in scene size and type. ScanNet scenes are often small-sized single-rooms with low scene diversity. Scenes in MP3D and HM3D are often large-sized multi-room settings with high diversity. MP3D also contains outdoor scenes. See figure S2 for a comparison of the evaluation datasets.

Training Trajectories. To sample the training set trajectories, for ScanNet and ScanNet++, we employ two sets of thresholds: $(t_{\min}, t_{\max}) \in \{(30\%, 70\%), (30\%, 100\%)\}$. For each set of thresholds, we sample $1k$ trajectories of 10 views per scene, for a total of $3.16M$ trajectories from ScanNet and ScanNet++. For HM3D and Gibson, we only deploy the more challenging threshold set $(t_{\min}, t_{\max}) = (30\%, 70\%)$ because we empirically observe overfitting in the experiments using the other threshold choice. Here, we sample 6K trajectories per scene with 10 views each, for a total of 7.8M trajectories. Note that we sample more trajectories because HM3D and Gibson scenes (Ramakrishnan et al., 2021) cover a much larger area than those in ScanNet and ScanNet++.

To generate a trajectory, we first choose as a starting view either one random view from the predefined trajectories in the ScanNet and ScanNet++ dataset, or render from a random location in Habitat-Sim (Savva et al., 2019) for the HM3D, Gibson, and MP3D datasets. Then we iteratively sample a new random view and add it to the trajectory if it meets a criterion (described later). We repeat this step until a total of 10 views are sampled. We now describe whether a new candidate view I^i meets the criterion. Let's refer to its pointmap as X^i and let $\{X^j\}_{j < i}$ denote the pointmaps of views that are already sampled. Using those, we calculate the overlap ratio $O(X^i, X^j)$ of each pair (X^i, X^j) . Formally,

$$O(X^i, X^j) = \frac{1}{2}(\text{Cov}(X^i, X^j) + \text{Cov}(X^j, X^i)), \quad (\text{S7a})$$

$$\text{Cov}(X^i, X^j) = \frac{1}{|A|} \sum_{p \in X^i} [\text{NearestDis}(p, X^j) < t_c], \quad (\text{S7b})$$

where $|A|$ denotes the number of points in X_i and $t_c = 0.0015$ is the threshold on the 3D point distance. A new view X^i is selected if the largest overlap ratio $\max_j \{O(X^i, X^j)\}_j$ is between t_{\min} and t_{\max} .

Test Trajectories. For testing, we generate 1K trajectories for each dataset following the procedure described above. The only difference: we sample 30-view trajectories, including 24 input views and 6 novel views for NVS evaluation. Those 6 novel views are selected to make sure that the 1st novel view is covered well by input views 1~4, the 2nd novel view is covered well by input views 5~8 and so on. To evaluate NVS with 4-view input, we use the extra 1st view for evaluation; for 8-view input, we use the extra 1st and 2nd view for evaluation and so on.

C.2 Multi-View Stereo Reconstruction

More results of up to 24 input views. In [table S2](#), we present additional results for input views $N \in \{8, 16, 20\}$. Those results are not presented in the main paper’s Table 2. Consistent with the findings in Table 2, Spann3R struggles in all settings on all 3 evaluation datasets. MV-DUSt3R+ consistently outperforms all other approaches in all settings, while MV-DUSt3R substantially outperforms DUSt3R by a margin.

Results of 100-view input. Next we validate the generalization performance of MV-DUSt3R+ using 100 input views while the number of input views at training time is fixed to 8. Specifically, we present qualitative MVS reconstruction results on ScanNet (see [figure S3](#) and [figure S4](#)) and HM3D (see [figure S5](#) and [figure S6](#)). For ScanNet data, we uniformly sample 100 views from the pre-defined trajectory in the original dataset. For HM3D data, we adopt the trajectory generation approach described in Section 4.1 of the main paper, but use less restrictive view overlapping thresholds $(t_{\min}, t_{\max}) = (50\%, 100\%)$ to sample a much larger number of input views.

As shown in [figure S3](#) and [figure S4](#), MV-DUSt3R+ generalizes to 100-view input remarkably well on ScanNet, and is able to reconstruct a scene geometry close to the groundtruth. On HM3D data, which contains much larger multi-room scenes, our MV-DUSt3R+ still performs remarkably well in reconstructing the scene (e.g., layout, walls, ceiling, objects), as shown in [figure S5](#) and [figure S6](#).

C.3 Multi-View Pose Estimation

Camera Intrinsics Estimation. We assume the camera intrinsics are the same for all the views in the trajectory, and thus only estimate it for the 1st view. At inference time, both MV-DUSt3R and MV-DUSt3R+ predict pointmap $\bar{X}^{1,1}$ for the 1st reference view I^1 in its own camera coordinate system. We further assumes the principal point is in the center of the image plane, and a pixel is square. The remaining unknown camera intrinsic parameter f_1^* can be estimated by solving the following minimization problem:

$$f_1^* = \arg \min_{f_1} \sum_{(i,j)} C_{i,j}^{1,1} \left\| (i', j') - f_1 \frac{(\bar{X}_{i,j,0}^{1,1}, \bar{X}_{i,j,1}^{1,1})}{\bar{X}_{i,j,2}^{1,1}} \right\|_2^2. \quad (\text{S8})$$

Here, $(i', j') = (i - \frac{H}{2}, j - \frac{W}{2})$ denotes the normalized 2D image coordinate. Following DUSt3R, we use the fast Weiszfeld algorithm ([Plastria, 2011](#)) to estimate f_1^* .

Camera Pose Estimation. To estimate the relative camera pose between 2 input views I^i and I^j , we first estimate per-view camera pose P_k^* for $k \in \{1, \dots, N\}$, where $P^k \in \mathbb{R}^{4 \times 4}$. For this we use RANSAC ([Fischler and Bolles, 1981](#)) with PnP ([Lepetit et al., 2009](#)) based on the correspondences between 2D pixels and the corresponding 3D pointmap $\bar{X}^{k,1}$, predicted by MV-DUSt3R or MV-DUSt3R+. In practice, for the 1st view I^1 , the estimated P^1 is often slightly different from the ideal identity matrix. After that, we estimate the relative camera pose as $P^j(P^i)^{-1}$.

Additional Results. In [table S3](#), we report MVPE results for different numbers of input views. Consistent with the findings in Table 3 of the main paper, MV-DUSt3R+ performs best in all settings on all 3 evaluation datasets, and MV-DUSt3R consistently outperforms DUSt3R.

We also report results of PoseDiffusion ([Wang et al., 2023b](#)) on HM3D and ScanNet. We use their Github open-source implementation² and load the checkpoint of the model trained on RealEstate10K data ([Zhou](#)

²<https://github.com/facebookresearch/PoseDiffusion>

et al., 2018b). As shown in [table S4](#), PoseDiffusion struggles on HM3D and ScanNet. We hypothesize that this is because PoseDiffusion is trained on input views with only slightly different camera poses and less diverse indoor scenes (see visualization in the original paper ([Wang et al., 2023b](#))). It thus is not able to generalize to our more challenging HM3D and ScanNet data, where input views are more sparse and the scenes are larger and more diverse.

C.4 Novel View Synthesis

DUSt3R baseline. The DUSt3R-based baseline uses manually defined per-pixel Gaussian parameters. We use the pointmap predicted by DUSt3R as the Gaussian center, use pixel RGB color I^v as the color, a constant 0.001 for the scale factor $S^{v,m}$, an identity transform, 1.0 for opacity, and spherical harmonics with degree zero. The scale constant is carefully tuned to balance between holes and blurry results. We also remove 3% of the lower confidence points to make the rendering quality as compelling as possible. This is a representative baseline given many pose-free reconstruction methods ([Fan et al., 2024](#); [Liu et al., 2024a](#)) are based on DUSt3R with global optimization. To render new views in a relative-scale world space, we follow [appendix A](#) in scaling the scene.

Additional Results. In [table S5](#), we report additional NVS results varying the number of input views, including 8, 16, and 20 views, which are not reported in Table 4 of the main paper. As more views are used and the scene gets larger, the performance of all methods decreases. However, MV-DUSt3R+ performs the best in all settings on all 3 evaluation datasets, while MV-DUSt3R consistently outperforms the DUSt3R based baseline by a large margin but moderately underperforms MV-DUSt3R+.

C.5 Study: MV-DUSt3R+ 2-Stage Training

Below we present more ablative studies and a discussion comparing 1-stage training with fixed 8-view inputs to 2-stage training using a mixed set of inputs with differing numbers of views. As in Section 4.7 of the main paper, in 2-stage training, we finetune the model trained in the 1st stage with fixed 8-view input for another 70 epochs. For this, we randomly sample inputs with a varying number of views N sampled between 4 and 12 views. In [table S6](#), we find the models with 2-stage training consistently outperform the models trained with 1-stage training on datasets of large scenes, including HM3D and MP3D. Models with 2-stage training perform comparably on small scenes in ScanNet. The improvements on MP3D are more significant when the number of input views is large (e.g., 24 views). We conclude, the 2nd stage training with mixed input views forces the model to be more robust to scene size and the number of input views. This improves the model’s generalization performance on large scenes.

	Method	GO	HM3D			ScanNet			MP3D			Time (sec)
			ND ↓	DAc ↑	CD ↓	ND ↓	DAc ↑	CD ↓	ND ↓	DAc ↑	CD ↓	
4 views	Spann3R	✗	37.1	0.0	225(184)	8.9	19.5	54.7(50.1)	42.7	0.0	248(202)	0.36
	DUSt3R	✓	1.9	75.1	5.6(2.3)	1.3	89.8	4.0(0.4)	3.9	41.7	40.0(5.3)	2.42
	MV-DUSt3R	✗	1.1	92.2	2.0(1.1)	1.0	93.3	2.0(0.4)	2.5	62.4	25.3(4.1)	0.05
	MV-DUSt3R+	✗	1.0	95.2	1.5(0.9)	0.8	94.9	1.5(0.3)	2.2	68.0	19.9(3.4)	0.29
	MV-DUSt3R _{oracle}	✗	1.0	94.6	1.5(0.7)	0.8	95.5	1.3(0.3)	2.3	66.6	20.7(4.0)	-
	MV-DUSt3R+ _{oracle}	✗	0.9	96.5	1.4(0.7)	0.7	95.8	1.2(0.2)	2.1	70.6	17.9(3.3)	-
8 views	Spann3R	✗	37.1	0.0	225(184)	8.9	19.5	54.7(50.1)	42.7	0.0	248(202)	0.79
	DUSt3R	✓	3.0	48.5	11.9(2.7)	1.7	86.4	4.2(0.6)	5.1	21.5	40.3(5.7)	4.49
	MV-DUSt3R	✗	1.4	87.4	2.5(1.2)	1.2	90.4	2.0(0.4)	2.9	53.8	21.1(5.1)	0.10
	MV-DUSt3R+	✗	1.1	93.5	2.9(0.8)	1.0	92.1	1.5(0.4)	2.3	62.0	15.1(4.2)	0.56
	MV-DUSt3R _{oracle}	✗	1.1	92.4	1.7(1.1)	0.9	92.8	1.3(0.4)	2.5	59.4	16.7(4.0)	-
	MV-DUSt3R+ _{oracle}	✗	1.0	95.3	1.8(0.8)	0.9	93.2	1.3(0.4)	2.2	65.4	14.0(4.2)	-
12 views	Spann3R	✗	32.6	0.0	125(113)	9.1	16.3	36.6(31.2)	35.0	0.0	138(112)	1.34
	DUSt3R	✓	3.9	30.7	18.1(3.4)	1.9	82.6	4.1(0.6)	6.6	12.0	49.6(8.3)	8.28
	MV-DUSt3R	✗	1.6	79.5	3.0(1.2)	1.4	86.8	2.3(0.8)	3.4	41.3	22.6(5.5)	0.15
	MV-DUSt3R+	✗	1.2	91.5	1.8(0.7)	1.2	88.4	1.8(0.7)	2.6	55.0	15.1(3.8)	0.89
	MV-DUSt3R _{oracle}	✗	1.3	88.8	1.8(0.9)	1.0	90.6	1.3(0.7)	2.9	51.3	16.4(4.0)	-
	MV-DUSt3R+ _{oracle}	✗	1.1	94.8	1.4(0.7)	1.0	90.9	1.3(0.5)	2.5	59.8	13.6(3.5)	-
16 views	Spann3R	✗	37.1	0.0	225(184)	8.9	19.5	54.7(50.1)	42.7	0.0	248(202)	1.73
	DUSt3R	✓	4.8	19.8	21.4(4.1)	2.1	78.1	3.9(0.9)	8.1	7.1	60.5(11.1)	13.07
	MV-DUSt3R	✗	2.0	67.5	4.2(2.1)	1.7	84.3	2.3(1.3)	4.2	29.2	26.1(8.5)	0.21
	MV-DUSt3R+	✗	1.5	85.5	2.3(1.5)	1.4	86.6	1.8(0.8)	3.2	44.1	17.3(4.4)	1.19
	MV-DUSt3R _{oracle}	✗	1.5	82.5	2.1(1.5)	1.2	89.3	1.3(0.7)	3.3	39.6	17.4(4.5)	-
	MV-DUSt3R+ _{oracle}	✗	1.3	90.2	1.7(1.1)	1.2	89.6	1.3(0.7)	2.8	51.4	14.2(3.2)	-
20 views	Spann3R	✗	37.1	0.0	225(184)	8.9	19.5	54.7(50.1)	42.7	0.0	248(202)	2.19
	DUSt3R	✓	5.7	12.0	27.1(4.9)	2.3	74.3	4.6(0.9)	9.7	4.1	69.2(11.1)	19.59
	MV-DUSt3R	✗	2.7	51.6	6.7(2.5)	1.9	79.3	2.5(1.0)	5.2	19.8	33.1(10.5)	0.28
	MV-DUSt3R+	✗	1.8	76.9	3.1(1.2)	1.5	84.0	1.8(0.7)	3.6	34.2	18.1(6.7)	1.54
	MV-DUSt3R _{oracle}	✗	1.8	72.5	2.7(1.3)	1.3	86.4	1.4(0.6)	3.7	30.8	17.3(5.8)	-
	MV-DUSt3R+ _{oracle}	✗	1.5	84.3	2.1(1.0)	1.3	86.8	1.3(0.5)	3.2	41.2	13.0(4.8)	-
24 views	Spann3R	✗	41.7	0.0	139(121)	11.4	1.6	37.4(35.5)	46.6	0.0	151(121)	2.73
	DUSt3R	✓	6.8	7.3	32.4(5.2)	2.4	72.6	5.1(1.0)	11.4	2.5	80.9(14.3)	27.21
	MV-DUSt3R	✗	3.4	36.7	10.0(3.5)	2.2	75.2	2.7(0.9)	6.3	12.2	38.6(13.9)	0.35
	MV-DUSt3R+	✗	2.1	64.5	3.9(2.0)	1.6	81.2	1.7(0.7)	4.3	26.7	22.0(5.9)	1.97
	MV-DUSt3R _{oracle}	✗	2.1	58.9	3.5(2.1)	1.4	82.9	1.4(0.7)	4.4	22.0	19.9(5.1)	-
	MV-DUSt3R+ _{oracle}	✗	1.8	77.9	2.6(1.3)	1.3	85.1	1.3(0.6)	3.6	33.1	15.1(4.4)	-

Table S2 Additional MVS reconstruction results. We use the same notation and scales of metrics as Table 2 in the main paper (the same below).

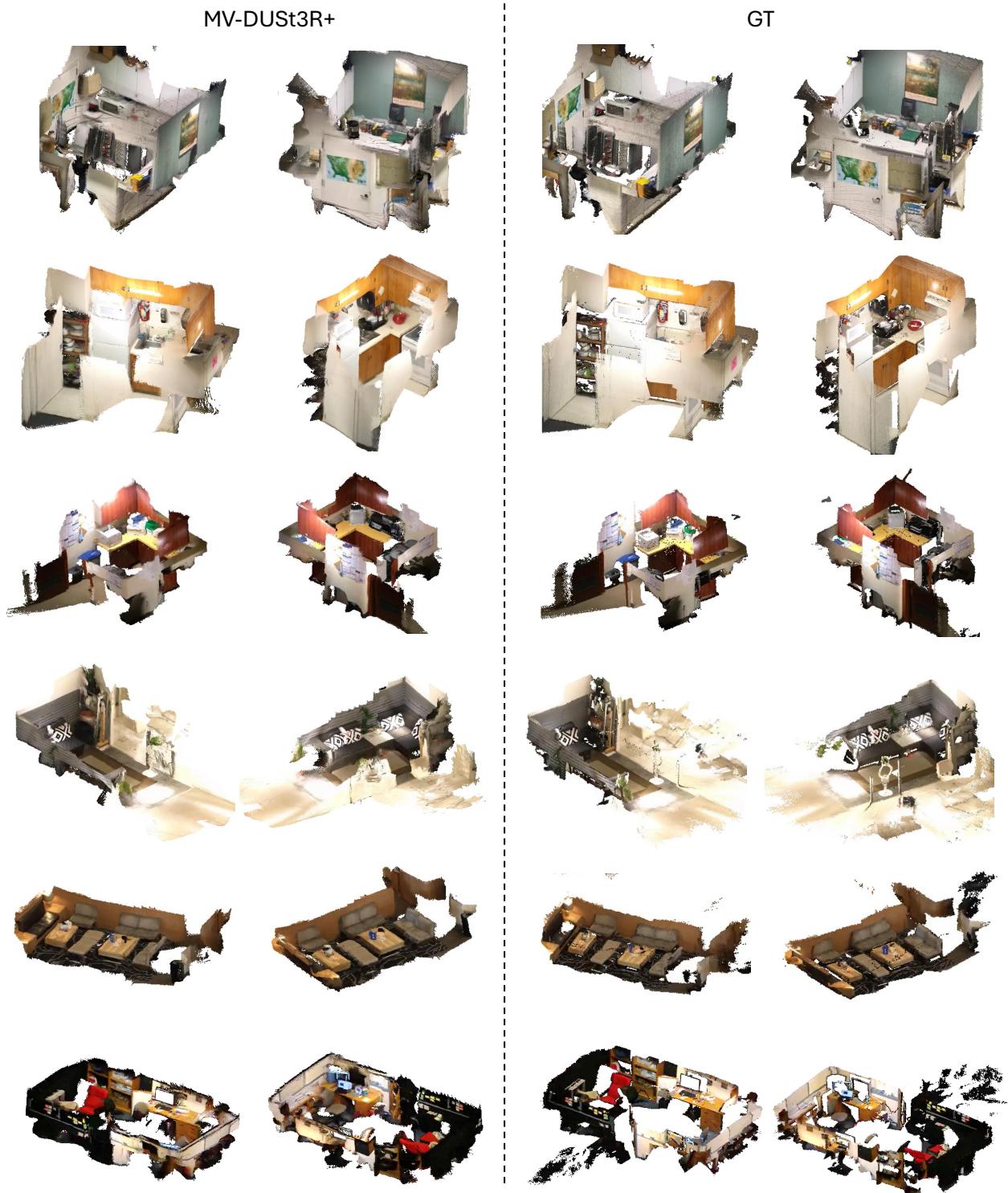


Figure S3 MVS reconstruction results with 100-view inputs on ScanNet. 3% of the predicted points with low confidence score are filtered out (the same below). Qualitatively we find the reconstructed scene geometry to be very similar to the groundtruth. This shows that our MV-DUSt3R+ model, trained with 8-view samples, can generalize remarkably well to 100-view inputs for single-room scenes. The inference time for 100 views is 19.1 seconds.

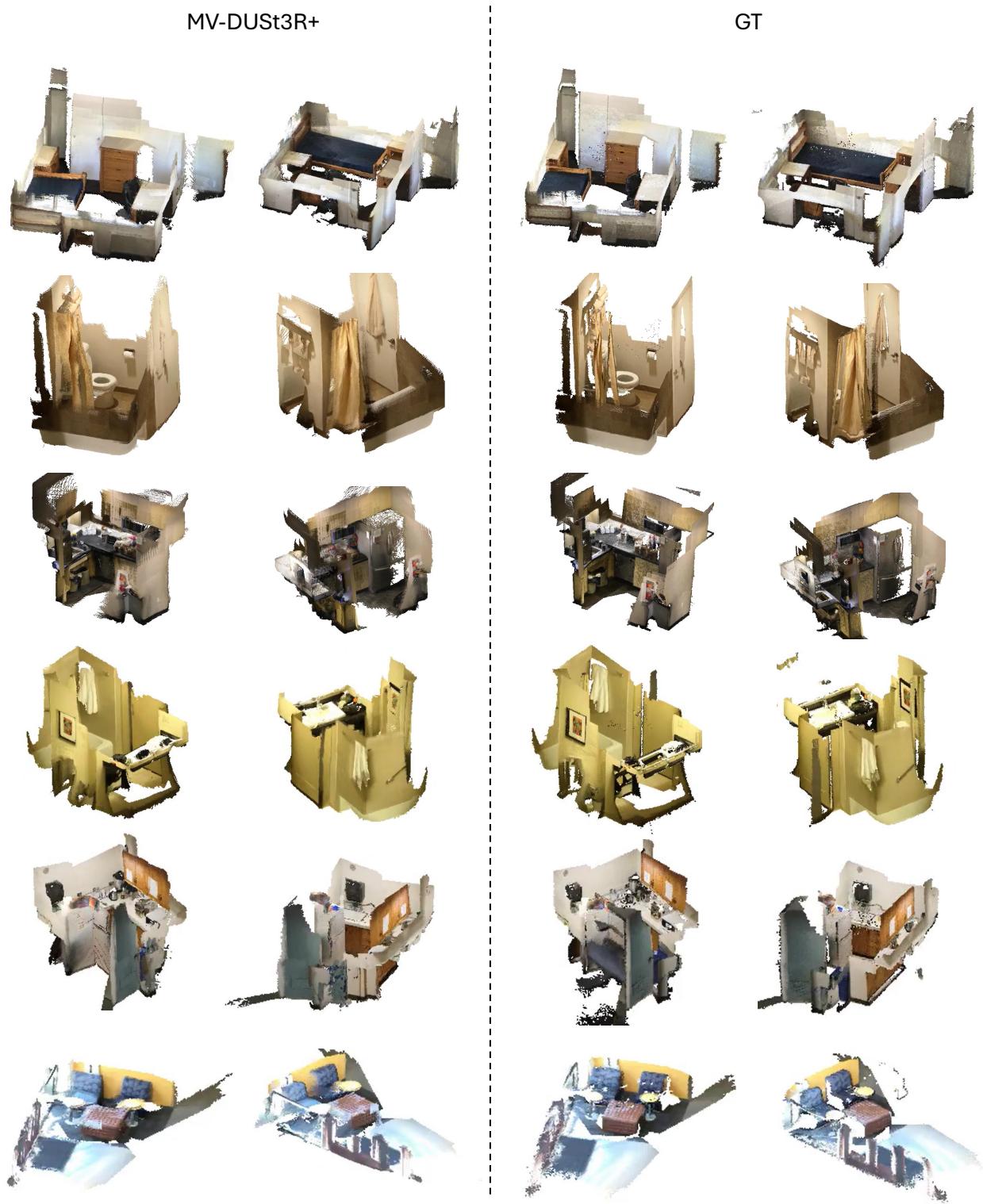


Figure S4 Extra MVS reconstruction results with 100-view inputs on ScanNet.

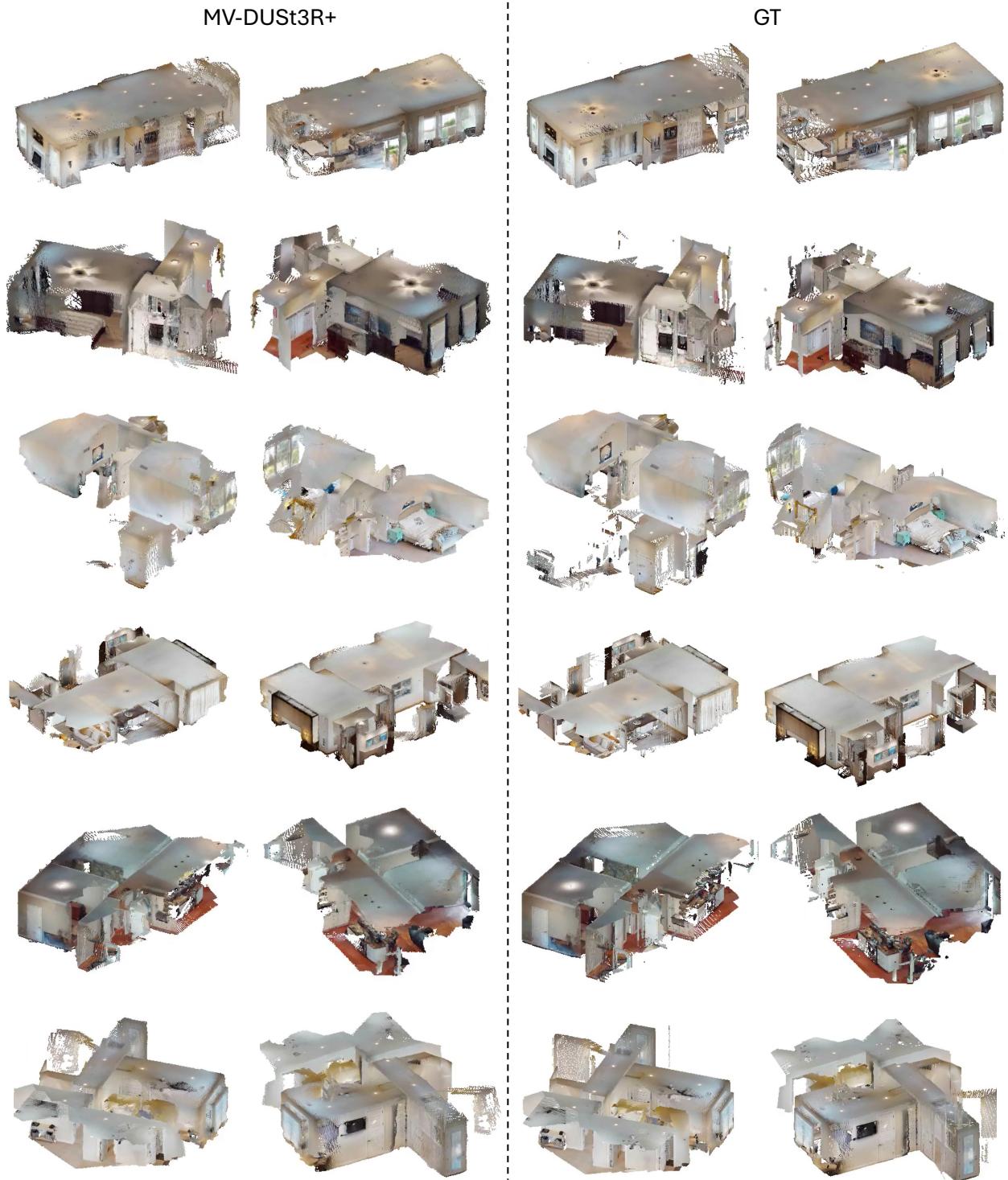


Figure S5 MVS reconstruction results with 100-view inputs on HM3D. MV-DUSt3R+ still performs well on more challenging multi-room scenes.

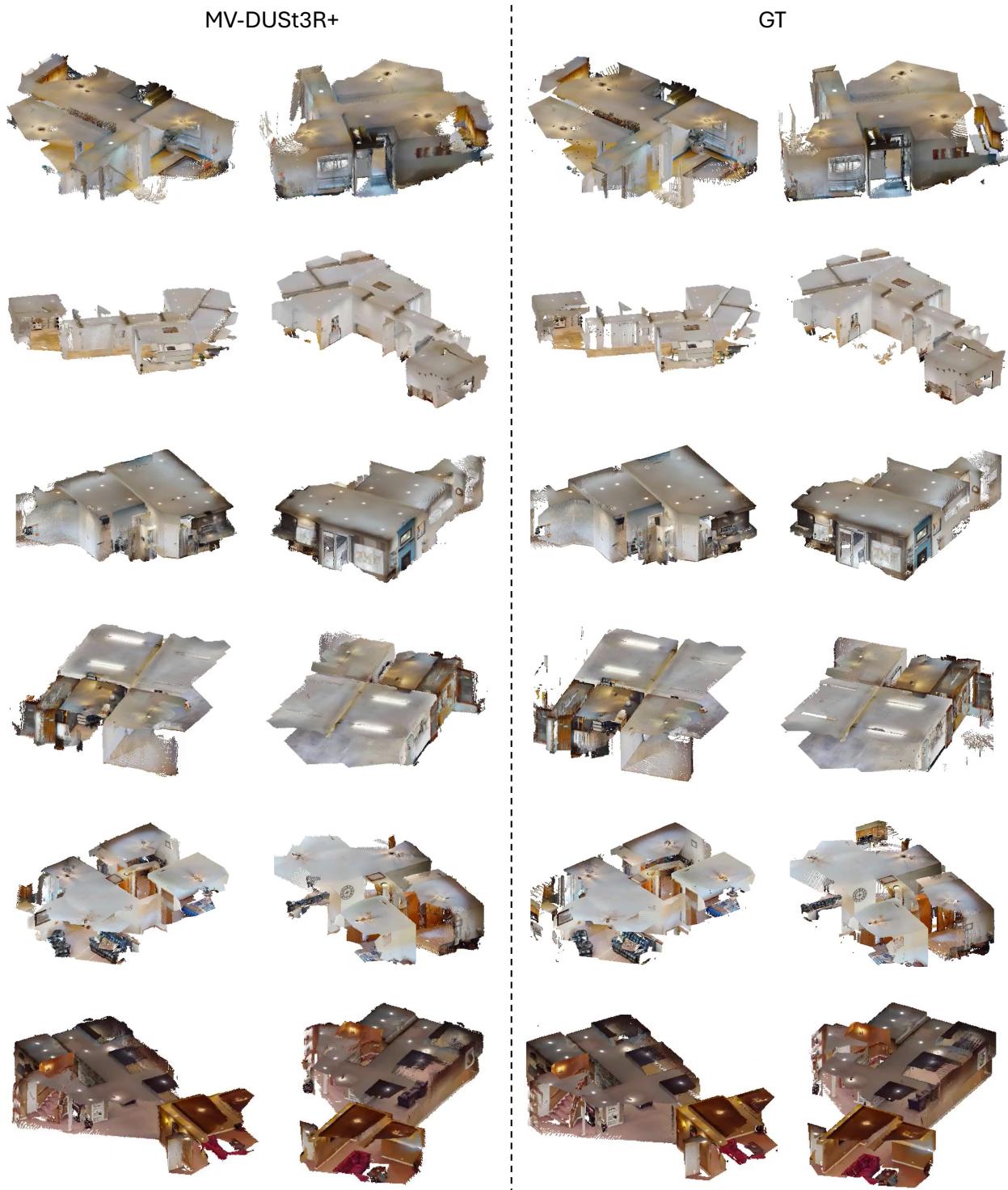


Figure S6 Additional MVS reconstruction results with 100-view inputs on HM3D.

	Method	GO	HM3D			ScanNet			MP3D		
			RRE ↓	RTE ↓	mAE ↓	RRE ↓	RTE ↓	mAE ↓	RRE ↓	RTE ↓	mAE ↓
4 views	DUSt3R	✓	2.4	3.1	12.5	3.0	20.0	30.7	3.5	3.8	13.3
	MV-DUSt3R	✗	1.5	1.5	5.5	2.3	16.8	27.0	1.2	1.0	5.4
	MV-DUSt3R+	✗	1.2	1.1	4.9	1.4	16.1	26.2	0.8	0.8	4.6
	MV-DUSt3R _{oracle}	✗	0.0	0.1	2.8	0.9	7.0	18.9	0.1	0.1	2.9
	MV-DUSt3R+ _{oracle}	✗	0.0	0.0	2.4	0.9	6.8	18.7	0.1	0.0	2.4
8 views	DUSt3R	✓	3.0	5.8	16.5	4.0	22.0	33.0	3.0	5.3	16.0
	MV-DUSt3R	✗	1.1	1.5	5.1	2.9	12.9	24.0	0.9	1.2	4.9
	MV-DUSt3R+	✗	0.6	0.8	2.9	2.0	10.1	21.0	0.4	0.5	2.6
	MV-DUSt3R _{oracle}	✗	0.1	0.2	2.8	1.6	6.5	18.7	0.1	0.2	2.8
	MV-DUSt3R+ _{oracle}	✗	0.1	0.1	1.6	1.4	5.3	16.6	0.1	0.1	1.5
12 views	DUSt3R	✓	3.7	8.3	20.1	4.6	22.6	34.2	4.5	8.4	19.8
	MV-DUSt3R	✗	1.5	2.6	8.4	3.7	14.7	26.1	1.6	2.6	8.2
	MV-DUSt3R+	✗	0.6	1.2	5.2	2.5	11.6	22.9	0.5	1.0	4.9
	MV-DUSt3R _{oracle}	✗	0.4	0.6	4.9	1.7	7.8	20.2	0.6	0.7	5.1
	MV-DUSt3R+ _{oracle}	✗	0.3	0.3	3.4	1.7	6.0	17.9	0.3	0.3	3.3
16 views	DUSt3R	✓	5.0	11.3	23.6	6.3	24.6	36.3	6.1	11.5	23.6
	MV-DUSt3R	✗	3.0	5.2	13.0	4.9	17.2	28.9	3.1	5.0	12.5
	MV-DUSt3R+	✗	1.0	2.5	8.5	3.1	13.4	25.3	1.4	2.4	8.1
	MV-DUSt3R _{oracle}	✗	0.7	1.1	7.5	2.4	9.8	22.7	1.1	1.3	7.8
	MV-DUSt3R+ _{oracle}	✗	0.5	0.6	5.7	2.0	7.6	20.3	0.8	0.8	5.7
20 views	DUSt3R	✓	6.5	14.6	27.2	7.7	26.0	38.1	7.9	14.8	27.1
	MV-DUSt3R	✗	5.5	8.7	18.1	6.6	19.7	31.7	5.1	7.8	16.7
	MV-DUSt3R+	✗	1.6	4.4	12.2	3.8	15.2	27.6	1.9	3.8	11.1
	MV-DUSt3R _{oracle}	✗	1.7	2.5	10.8	2.7	11.4	24.8	1.8	2.5	10.5
	MV-DUSt3R+ _{oracle}	✗	0.7	1.3	8.3	2.3	8.7	22.2	1.0	1.4	7.9
24 views	DUSt3R	✓	8.8	18.1	30.9	8.1	26.6	38.9	10.0	18.2	30.5
	MV-DUSt3R	✗	8.9	12.8	23.7	8.2	21.9	34.2	8.2	11.1	21.4
	MV-DUSt3R+	✗	3.0	6.5	15.8	4.6	16.7	29.4	3.3	6.0	14.6
	MV-DUSt3R _{oracle}	✗	3.2	4.4	14.7	3.4	13.1	26.7	3.4	4.2	14.0
	MV-DUSt3R+ _{oracle}	✗	1.4	2.4	11.1	2.6	9.9	23.7	1.8	2.4	10.6

Table S3 Additional Multi-View Pose Estimation results.

# of views	HM3D			ScanNet		
	RRE ↓	RTE ↓	mAE ↓	RRE ↓	RTE ↓	mAE ↓
4	93.3	90.0	100.0	96.7	86.7	98.9
12	94.5	94.8	98.8	97.0	98.5	99.9
24	95.1	96.7	99.5	98.5	96.3	99.8

Table S4 PoseDiffusion Evaluation Results. Metrics are reported in percent %.

	Method	GO	HM3D			ScanNet			MP3D		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
4 views	DUS3R	✓	16.0	5.0	3.7	17.0	6.0	3.0	15.5	4.6	4.0
	MV-DUSt3R	✗	19.9	6.0	2.0	21.9	7.1	1.6	19.6	5.8	2.1
	MV-DUSt3R+	✗	20.2	6.1	1.9	22.2	7.1	1.5	19.9	5.9	2.0
	MV-DUSt3R _{oracle}	✗	21.0	6.5	1.6	22.8	7.4	1.4	20.6	6.2	1.8
	MV-DUSt3R+ _{oracle}	✗	21.4	6.6	1.5	23.0	7.4	1.4	21.0	6.3	1.7
	DUS3R	✓	15.5	4.6	4.4	16.7	5.6	3.4	15.0	4.2	4.8
8 views	MV-DUSt3R	✗	19.3	5.8	2.3	21.0	6.8	1.9	18.9	5.5	2.5
	MV-DUSt3R+	✗	19.8	6.0	2.1	21.3	6.9	1.8	19.5	5.7	2.2
	MV-DUSt3R _{oracle}	✗	20.6	6.2	1.9	21.9	7.1	1.6	20.0	6.0	2.0
	MV-DUSt3R+ _{oracle}	✗	21.2	6.5	1.7	22.2	7.2	1.6	20.7	6.2	1.8
	DUS3R	✓	15.1	4.4	4.9	16.3	5.4	3.6	14.7	4.0	5.3
	MV-DUSt3R	✗	18.9	5.6	2.7	20.1	6.5	2.2	18.4	5.3	2.8
12 views	MV-DUSt3R+	✗	19.4	5.8	2.4	20.4	6.6	2.1	19.0	5.5	2.6
	MV-DUSt3R _{oracle}	✗	19.9	5.9	2.2	21.0	6.8	1.9	19.3	5.6	2.4
	MV-DUSt3R+ _{oracle}	✗	20.4	6.1	2.0	21.3	6.8	1.8	19.9	5.8	2.2
	DUS3R	✓	14.8	4.3	5.2	15.8	5.2	3.9	14.3	3.8	5.6
	MV-DUSt3R	✗	18.5	5.5	3.0	19.4	6.3	2.4	18.0	5.2	3.2
	MV-DUSt3R+	✗	19.0	5.6	2.7	19.6	6.3	2.3	18.6	5.3	2.9
16 views	MV-DUSt3R _{oracle}	✗	19.3	5.8	2.5	20.4	6.6	2.1	18.8	5.4	2.7
	MV-DUSt3R+ _{oracle}	✗	19.8	5.9	2.3	20.5	6.6	2.1	19.4	5.6	2.5
	DUS3R	✓	14.6	4.2	5.4	15.5	5.1	4.1	14.1	3.7	5.9
	MV-DUSt3R	✗	18.1	5.4	3.3	18.8	6.1	2.7	17.7	5.0	63.5
	MV-DUSt3R+	✗	18.6	5.5	3.0	19.0	6.1	2.6	18.3	5.2	3.1
	MV-DUSt3R _{oracle}	✗	18.9	5.6	2.8	19.8	6.4	2.3	18.4	5.3	3.0
20 views	MV-DUSt3R+ _{oracle}	✗	19.4	5.7	2.6	19.9	6.4	2.3	18.9	5.4	2.8
	DUS3R	✓	14.3	4.2	5.6	15.2	5.0	4.2	13.8	3.6	6.1
	MV-DUSt3R	✗	17.8	5.3	3.6	18.4	6.0	2.9	17.3	4.9	3.8
	MV-DUSt3R+	✗	18.4	5.4	3.2	18.6	6.0	2.8	17.9	5.1	3.5
	MV-DUSt3R _{oracle}	✗	18.5	5.5	3.1	19.3	6.2	2.5	18.0	5.1	3.3
	MV-DUSt3R+ _{oracle}	✗	19.0	5.6	2.9	19.4	6.2	2.5	18.5	5.3	3.1
24 views	DUS3R	✓	14.0	4.2	5.7	15.0	5.0	4.3	13.5	3.6	6.3
	MV-DUSt3R	✗	17.5	5.3	3.5	18.3	6.0	2.9	17.0	4.9	3.8
	MV-DUSt3R+	✗	18.0	5.4	3.1	18.5	6.0	2.8	17.6	5.1	3.5
	MV-DUSt3R _{oracle}	✗	18.5	5.5	3.0	19.3	6.2	2.5	18.0	5.1	3.3
	MV-DUSt3R+ _{oracle}	✗	19.0	5.6	2.9	19.4	6.2	2.5	18.5	5.3	3.1

Table S5 Additional Novel View Synthesis results.

Dataset	# of views	Training recipe	MVS Reconstruction			MVPE			NVS		
			ND \downarrow	DAc \uparrow	CD \downarrow	RRE \downarrow	RTE \downarrow	mAE \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
HM3D	4	1-stage, 8 views	1.0	95.2	1.5(0.9)	1.2	1.1	4.9	20.2	6.1	1.9
		2-stage, mixed views	0.9	95.5	1.5(0.7)	0.7	0.6	1.9	20.7	6.3	1.8
	8	1-stage, 8 views	1.1	93.5	2.9(0.8)	0.6	0.8	2.9	19.8	6.0	2.1
		2-stage, mixed views	1.1	94.0	2.0(1.3)	0.5	0.7	2.6	19.9	6.0	2.1
	12	1-stage, 8 views	1.2	91.5	1.8(0.7)	0.6	1.2	5.2	19.4	5.8	2.4
		2-stage, mixed views	1.1	93.2	1.5(1.0)	0.4	0.7	3.6	19.5	5.9	2.2
	16	1-stage, 8 views	1.5	85.5	2.3(1.5)	1.0	2.5	8.5	19.0	5.6	2.7
		2-stage, mixed views	1.3	89.6	2.3(1.3)	0.6	1.1	5.0	19.4	5.8	2.4
	20	1-stage, 8 views	1.8	76.9	3.1(1.2)	1.6	4.4	12.2	18.6	5.5	3.0
		2-stage, mixed views	1.4	87.0	2.1(1.3)	0.8	1.9	6.7	19.2	5.8	2.5
	24	1-stage, 8 views	2.1	64.5	3.9(2.0)	3.0	6.5	15.8	18.4	5.4	3.2
		2-stage, mixed views	1.6	83.5	2.4(1.2)	1.2	2.7	8.5	19.1	5.7	2.6
ScanNet	4	1-stage, 8 views	0.8	94.9	1.5(0.3)	1.4	16.1	26.2	22.2	7.1	1.5
		2-stage, mixed views	0.9	94.5	1.8(0.3)	1.5	10.6	20.7	23.1	7.3	1.4
	8	1-stage, 8 views	1.0	92.1	1.5(0.4)	2.0	10.1	21.0	21.3	6.9	1.8
		2-stage, mixed views	1.1	90.2	1.8(0.5)	2.2	10.8	21.6	21.3	6.9	1.8
	12	1-stage, 8 views	1.2	88.4	1.8(0.7)	2.5	11.6	22.9	20.4	6.6	2.1
		2-stage, mixed views	1.2	87.6	1.7(0.8)	2.4	11.4	22.7	20.3	6.6	2.0
	16	1-stage, 8 views	1.4	86.6	1.8(0.8)	3.1	13.4	25.3	19.6	6.3	2.3
		2-stage, mixed views	1.4	85.7	2.0(0.6)	3.0	12.5	24.3	19.6	6.4	2.3
	20	1-stage, 8 views	1.5	84.0	1.8(0.7)	3.8	15.2	27.6	19.0	6.1	2.6
		2-stage, mixed views	1.5	82.9	2.0(0.7)	3.8	13.9	26.0	19.1	6.2	2.5
	24	1-stage, 8 views	1.6	81.2	1.7(0.7)	4.6	16.7	29.4	18.6	6.0	2.8
		2-stage, mixed views	1.5	81.5	1.8(0.6)	4.5	14.5	27.0	18.7	6.1	2.6
MP3D	4	1-stage, 8 views	2.2	68.0	19.9(3.4)	0.8	0.8	4.6	19.9	5.9	2.0
		2-stage, mixed views	2.1	70.4	19.9(3.1)	0.8	0.6	2.2	20.4	6.0	1.9
	8	1-stage, 8 views	2.3	62.0	15.1(4.2)	0.4	0.5	2.6	19.5	5.7	2.2
		2-stage, mixed views	2.3	63.4	14.9(3.4)	0.4	0.5	2.6	19.5	5.8	2.2
	12	1-stage, 8 views	2.6	55.0	15.1(3.8)	0.5	1.0	4.9	19.0	5.5	2.6
		2-stage, mixed views	2.5	57.2	14.0(2.7)	0.5	0.7	3.7	19.1	5.6	2.4
	16	1-stage, 8 views	3.2	44.1	17.3(4.4)	1.4	2.4	8.1	18.6	5.3	2.9
		2-stage, mixed views	2.8	52.0	14.0(3.6)	0.8	1.4	5.2	18.9	5.5	2.5
	20	1-stage, 8 views	3.6	34.2	18.1(6.7)	1.9	3.8	11.1	18.3	5.2	3.1
		2-stage, mixed views	3.1	44.8	14.1(3.2)	1.0	2.0	6.7	18.7	5.5	2.7
	24	1-stage, 8 views	4.3	26.7	22.0(5.9)	3.3	6.0	14.6	17.9	5.1	3.5
		2-stage, mixed views	3.4	38.1	16.1(3.0)	1.5	2.8	8.3	18.5	5.4	2.9

Table S6 Comparisons between 1-stage training and 2-stage training. Models trained with 2 stages perform substantially better on large scenes from HM3D and MP3D, while they perform similarly on small scenes from ScanNet.

References

- Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016.
- Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. *arXiv preprint arXiv:2403.13064*, 2024.
- Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.
- Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *CVPR*, 2023.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.
- Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.
- Sylvain Bougnoux. From projective to euclidean space under any practical situation, a criticism of self-calibration. In *ICCV*, 1998.
- Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *CVPR*, 2017.
- Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 2016.
- Jonathan L Carrivick, Mark W Smith, and Duncan J Quincey. *Structure from Motion in the Geosciences*. John Wiley & Sons, 2016.
- Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022.
- 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.
- David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.
- Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 2021.
- 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 *ECCV*, 2024.
- Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Neurips*, 2020.
- Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.
- P Kingma Diederik. Adam: A method for stochastic optimization. (*No Title*), 2014.
- Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *CVPR*, 2023.
- Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024.

- Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantssplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2, 2024.
- Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981.
- Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Neurips*, 2022.
- Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2015.
- Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015.
- Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020.
- Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.
- Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *CVPR*, 2024.
- Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In *CVPR*, 2024.
- Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024.
- Jakob Iglhaut, Carlos Cabo, Stefano Piliti, Livia Piermattei, James O'Connor, and Jacqueline Rosette. Structure from motion photogrammetry in forestry: A review. *Current Forestry Reports*, 2019.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023.
- Mustafa Khan, Hamidreza Fazlali, Dhruv Sharma, Tongtong Cao, Dongfeng Bai, Yuan Ren, and Bingbing Liu. Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction. *arXiv preprint arXiv:2407.02598*, 2024.
- Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *CVPR*, 2022.
- Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *IJCV*, 2009.
- Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024.
- Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024a.
- Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*, 2020.
- Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024b.
- David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
- Baorui Ma, Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Towards better gradient consistency for neural signed distance functions via level set alignment. In *CVPR*, 2023.
- 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*, 2021.

- Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMARISMAR*, 2011.
- Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 2017.
- Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Neurips*, 2019.
- MV Peppa, JP Mills, KD Fieber, I Haynes, S Turner, A Turner, M Douglas, and PG Bryan. Archaeological feature detection from archive aerial photography with a sfm-mvs and image enhancement pipeline. *ISPRS*, 2018.
- Frank Plastria. The weiszfeld algorithm: proof, amendments, and extensions. *Foundations of location analysis*, 2011.
- Long Quan and Zhongdan Lan. Linear n-point camera pose determination. *IEEE TPAMI*, 1999.
- Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.
- Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023.
- 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 *ICCV*, 2019.
- Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *ICCV*, 2023.
- Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.
- Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.
- Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Neurips*, 2019.
- Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Neurips*, 2020.
- 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.
- Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Zhao Yang. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. In *ACM MM*, 2023.
- Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, 2024.
- Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *ECCV*, pages 1–18. Springer, 2024.
- Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, 2000.
- Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *CVPR*, 2022.
- Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, 2021.

- Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.
- Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Visual geometry grounded deep structure from motion. *arXiv preprint arXiv:2312.04563*, 2023a.
- Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023b.
- Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *3DV*, 2018.
- Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.
- Yuesong Wang, Zhaojie Zeng, Tao Guan, Wei Yang, Zhuo Chen, Wenkai Liu, Luoyuan Xu, and Yawei Luo. Adaptive patch deformation for textureless-resilient multi-view stereo. In *CVPR*, 2023c.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004.
- Matthew J Westoby, James Brasington, Niel F Glasser, Michael J Hambrey, and Jennifer M Reynolds. ‘structure-from-motion’ photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 2012.
- Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024.
- Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024.
- Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, 2018.
- Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *ITSC*, 2021.
- Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *CVPR*, 2019.
- Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024.
- Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018.
- Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019.
- Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Neurips*, 2020.
- Botao Ye, Sifei Liu, Haofei Xu, Xuetong 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.
- Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.
- Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016.
- Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.
- Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024.
- Rui Zeng, Yuhui Wen, Wang Zhao, and Yong-Jin Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 2020.
- Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, pages 1–19. Springer, 2024.

- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.
- Zhengdong Zhang, Yasuyuki Matsushita, and Yi Ma. Camera calibration with lens distortion from low-rank textures. In *CVPR*, 2011.
- Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *CVPR*, 2014.
- Lei Zhou, Siyu Zhu, Zixin Luo, Tianwei Shen, Runze Zhang, Mingmin Zhen, Tian Fang, and Long Quan. Learning and matching multi-view descriptors for registration of point clouds. In *ECCV*, 2018a.
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018b.
- Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *CVPR*, 2024.