MoRE: 3D Visual Geometry Reconstruction Meets Mixture-of-Experts

Jingnan Gao[1,2*] Zhe Wang[2*] Xianze Fang[2] Xingyu Ren[1] Zhuo Chen[1] Shengqi Liu[1]
Yuhao Cheng[1] Jiangjing Lyu[2†] Xiaokang Yang[1] Yichao Yan[1‡]
[1]Shanghai Jiao Tong University [2]Alibaba Group

https://g-1nonly.github.io/MoRE_Website/

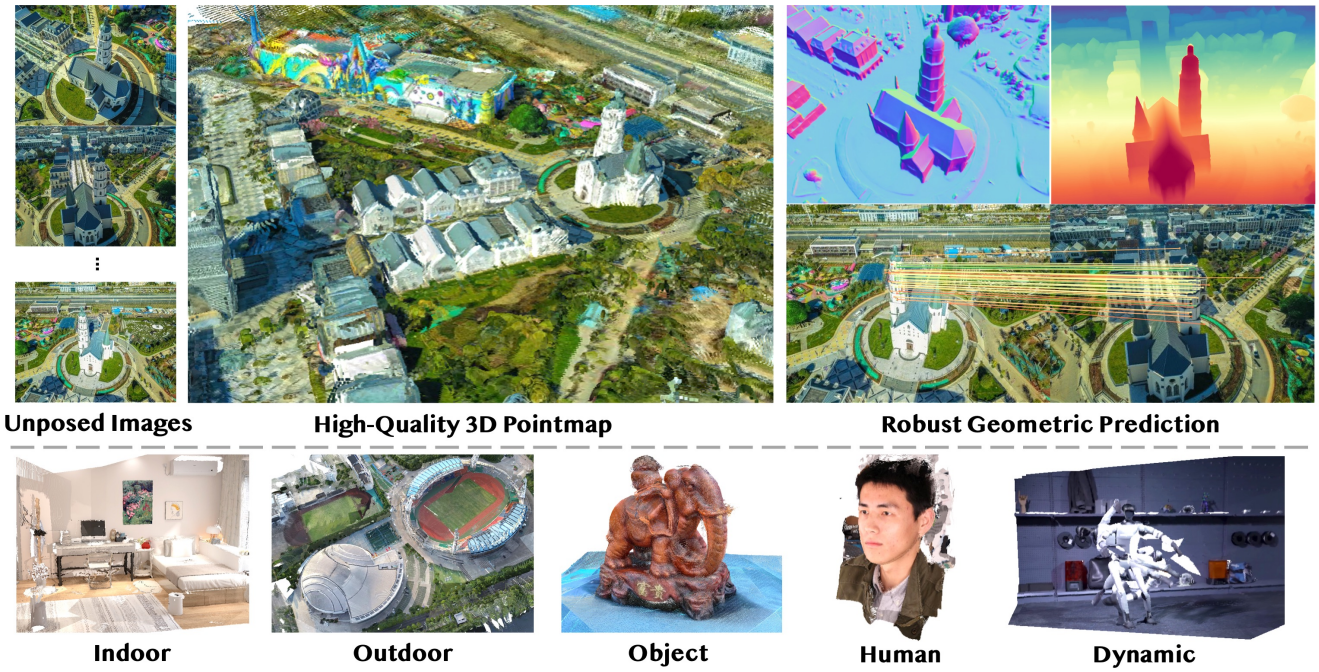

Figure 1. ***MoE sparks 3D visual geometry Reconstruction to do MoRE.*** MoRE is a feed-forward foundation model that leverages mixture-of-experts in 3D visual geometry reconstruction. MoRE takes unposed images as input and outputs high-quality 3D pointmap, achieving robust geometric predictions for various scenarios.

Abstract

Recent advances in language and vision have demonstrated that scaling up model capacity consistently improves performance across diverse tasks. In 3D visual geometry reconstruction, large-scale training has likewise proven effective for learning versatile representations. However, further scaling of 3D models is challenging due to the complexity of geometric supervision and the diversity of 3D data. To overcome these limitations, we propose MoRE, a dense 3D visual foundation model based on a Mixture-of-Experts (MoE) architecture that dynamically routes features to domain-specific experts, allowing them to specialize in complementary data aspects and enhance both scalability and adaptability. Aiming to improve robustness under real-world conditions, MoRE incorporates a confidence-based depth refinement module that stabilizes and refines geometric estimation. In addition, it integrates dense semantic features with globally aligned 3D backbone representations for high-fidelity surface normal prediction. MoRE is further optimized with tailored loss functions to ensure robust learning across diverse inputs and multiple geometric tasks. Extensive experiments demonstrate that MoRE achieves state-of-the-art performance across multiple benchmarks and supports effective downstream applications without extra computation.

1. Introduction

Traditional methods for 3D visual-geometry reconstruction typically rely on scene-specific optimization, where models are trained for each environment or dataset. Although such pipelines can achieve high accuracy in restricted set-

* Equal contribution † Project leader ‡ Corresponding author

tings, they lack the flexibility required by real-world applications that demand strong geometric priors and consistent performance across diverse scenes like AR/VR, game content creation, robotics, and autonomous driving. To overcome these limitations, researchers draw inspiration from recent foundation models such as GPTs [1, 72], CLIP [42], DINO [8, 38, 50], and Stable Diffusion [12], which demonstrate that scaling data and model capacity enables highly versatile representations across diverse tasks. Following this paradigm, the field of 3D visual geometry reconstruction is moving beyond task-specific optimization toward scalable and generalizable architectures. These models [7, 10, 13, 21, 29, 52, 54, 56–59, 68, 75, 76] demonstrate the potential of feed-forward networks to enable joint 3D geometric prediction across diverse input configurations.

Notably, a key factor behind the success of these models is large-scale training, with evidence showing that increasing model capacity consistently yields stronger results across domains. In the field of LLMs, scaling up Transformers has led to remarkable gains [1, 6, 20, 53], albeit at the cost of substantial computational resources. A similar scaling behavior is observed in vision, where the progression from DINOv2 [38] to DINOv3 [50] highlights the benefits of enlarging model capacity for representation learning. The same principle naturally extends to 3D visual geometry: training larger networks on larger-scale datasets can potentially improve accuracy across multiple 3D geometric prediction tasks. However, scaling 3D models is arguably even more demanding due to the complexity of geometric supervision and heterogeneous nature of 3D data.

To address these challenges, we draw inspiration from the Mixture-of-Experts (MoE) framework, which has proven effective in scaling neural networks [9, 11, 15, 28, 41, 65–67]. MoE activates only a subset of experts for each input, enabling model capacity to expand without a proportional increase in computation. This mechanism enables experts to specialize in complementary aspects of the data, improving both scalability and adaptability. Such specialization is particularly valuable for 3D geometry reconstruction, where scenes vary widely across indoor, outdoor, object-centric, human-centric, and dynamic environments. By exploiting expert specialization within a unified framework, MoE provides a principled way to scale 3D geometry models effectively while maintaining computational efficiency.

In this paper, we introduce **MoRE**, a large-scale 3D visual foundation model that unifies 3D visual geometry reconstruction with the Mixture-of-Experts paradigm. MoRE builds upon a dense visual transformer backbone and dynamically routes features to domain-specialized experts, enabling the model to learn adaptive and complementary representations for different 3D scenarios. To enhance geometric reliability, we incorporate a confidence-based depth refinement module that mitigates noise and inconsistency in real-world data. Furthermore, we fuse dense semantic features with globally aligned 3D representations to achieve more consistent and detailed surface normal estimation. Joint optimization across multiple 3D quantities is achieved through tailored loss functions that ensure stability and convergence during large-scale training. Extensive experiments demonstrate that our approach achieves highly accurate 3D reconstruction and sets new state-of-the-art results across multiple benchmarks.

In summary, we make the following contributions:

- We introduce **MoRE**, a dense 3D visual foundation model that leverages the **mixture-of-experts** framework in 3D geometric predictions and demonstrates high-quality performance in various 3D scenarios.
- We propose **dense semantic feature fusion** with **confidence-based depth refinement** to enhance the consistency and precision of 3D geometry estimation.
- Extensive experiments on various benchmarks showcase our **state-of-the-art** 3D geometric predictions.

2. Related Work

2.1. Feed-Forward 3D Reconstruction

With the advent of deep learning, feed-forward paradigms have emerged as a compelling alternative to traditional optimization-based pipelines for 3D reconstruction [3, 4, 17, 18, 26, 32, 34–37, 40, 45, 55, 61, 69, 70, 74, 77]. These methods exploit neural networks to encode strong scene priors, enabling direct regression of 3D structure from raw image inputs and improving robustness and generalization across datasets. Early progress was exemplified by DUSt3R [59], which produced pairwise-consistent point maps without calibration. However, its reliance on pairwise predictions necessitated a global alignment stage for larger scenes. Subsequent extensions introduced mechanisms to alleviate these limitations. MASt3R [29] incorporated confidence-weighted objectives to approximate metric scale, and Fast3R [68] scaled inference to thousands of views, thereby obviating alignment entirely. Other approaches aimed to reformulate the task, such as FLARE [76], which decouples pose estimation from geometry prediction. More recently, large-scale transformer models [7, 13, 21, 52, 54, 56, 57] like VGGT have advanced the state of the art by jointly predicting intrinsic and extrinsic parameters, dense depth, point maps, and feature correspondences. Subsequent approaches [10, 62, 63] further enhanced stability and accuracy through more sophisticated strategies. Despite these advances, existing feed-forward frameworks remain limited in scalability and generalization. Our approach tackles these issues by incorporating mixture-of-experts architecture, enabling more accurate and high-fidelity reconstruction.

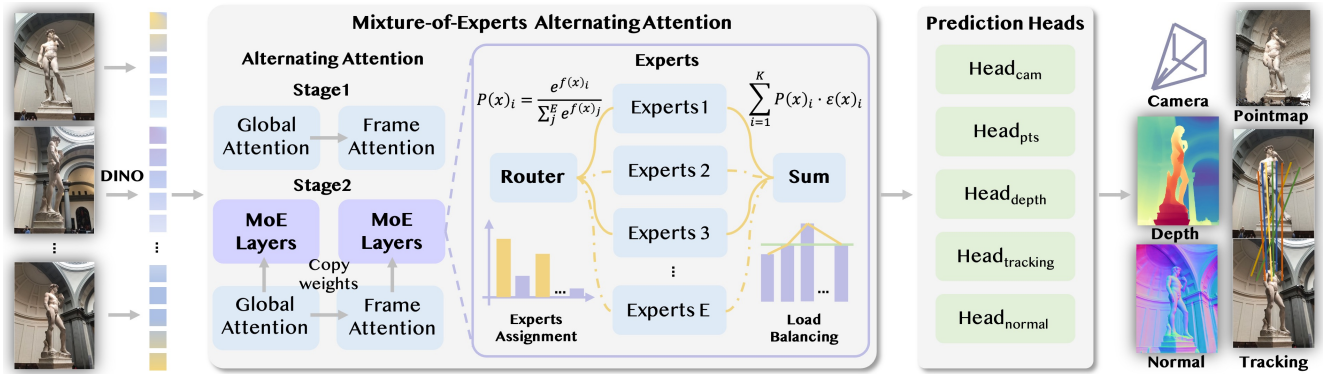

Figure 2. Overview of MoRE. We propose MoRE, a dense visual foundation model featuring a mixture-of-experts architecture and multiple task-specific heads for geometric prediction. We adopt a two-stage strategy during the model training. In Stage 1, we supervise our model with the multi-task training objectives. In Stage 2, we incorporate mixture-of-experts to further train the model for robust and accurate visual geometry reconstruction.

2.2. Mixture-of-Experts Framework

Recent large language models (LLMs) increasingly adopt the mixture-of-experts (MoE) framework to address the high computational cost of densely activated Transformers. As demonstrated by several early approaches [23, 47, 64], MoE achieves substantial efficiency gains while maintaining competitive performance by activating only a small subset of parameters during inference. Beyond efficiency, MoE [33, 60] has also been utilized for modality specialization, where different modalities are routed to dedicated experts. Several works [30, 31, 43, 79] have also explored converting pretrained dense LLMs into MoE models, combining the strengths of existing LLMs with the scalability and efficiency advantages of MoE. To further improve the performance across diverse downstream tasks, recent approaches [9, 65–67] have leveraged fine-grained expert segmentation and shared experts routing. Inspired by the effectiveness of MoE in improving prediction across diverse data domains, we propose a 3D foundation model built upon the MoE framework. Since 3D geometric reconstruction encompasses highly diverse data distributions, our approach leverages MoE to adaptively model such diversity and achieve robust geometric prediction.

3. Method

This work aims to build an end-to-end framework for predicting various 3D geometric quantities from unconstrained images input. Our approach consists of three main components. First, we employ a dense visual transformer backbone to extract features that adapt to unconstrained input requirements (Sec. 3.1). Second, we introduce a mixture-of-experts (MoE) mechanism into the geometric prediction pipeline, allowing the model to effectively enhance both accuracy and scalability and dynamically allocate its capacity across different scenarios (Sec. 3.2). Finally, we develop specialized training strategies that stabilize optimization and further improve the generalization of the learned representations to various tasks (Sec. 3.3). With these components, our proposed model **MoRE** can flexibly adapt to varying input scenarios while achieving strong performance across a wide range of 3D geometric vision problems.

3.1. Model Architecture

Inspired by recent progress in 3D vision [13, 24, 54, 62], our goal is to develop a foundation model that can predict a variety of geometric quantities across diverse scenes and tasks. To achieve this, we employ a dense visual transformer backbone trained on large-scale 3D annotated datasets. Given a sequence of N RGB images $(I_i)_{i=1}^N \in \mathcal{R}^{3\times H\times W}$, MoRE's dense visual transformer is a function f that maps the input to a corresponding set of 3D quantities per frame:

$$(C_i, P_i, D_i, T_i, N_i,)_{i=1}^N = f((I_i)_{i=1}^N), \tag{1}$$

where $C_i \in \mathcal{R}^9$ is the camera parameters including both intrinsics and extrinsics, $P_i \in \mathcal{R}^{3\times H\times W}$ is the pointmap, $D_i \in \mathcal{R}^{H\times W}$ is the depth map, $T_i \in \mathcal{R}^{C\times H\times W}$ is a grid of C-dimensional features for point tracking, and $N_i \in \mathcal{R}^{3\times H\times W}$ is the normal map. Apart from the pointmap head, depth head, camera head and tracking head implemented by VGGT [54], we additionally implement a normal prediction head to facilitate normal map estimation. We further design a confidence-based depth refinement for more accurate monocular depth estimation.

Confidence-based Depth Refinement. Real-world depth training data often contain noise and missing measurements, which can cause models to overfit unreliable depth ground truth and harm estimation accuracy. However, a state-of-the-art monocular model [58] with refined training data still produces reasonably accurate results as shown

Figure 3. Real-world depth comparison. We present the ground-truth depth, prediction from MoGe, the confidence mask and our prediction after training with confidence-based depth refinement.

in Fig. 3. To further leverage this insight, we design a confidence-based depth refinement to filter depth supervision. Specifically, we utilize MoGev2 [58] to compute the confidence masks for each depth sample:

$$M_{\text{conf}} = \left\{ \frac{|D_{\text{moge}} - D_{\text{gt}}|}{\max(D_{\text{gt}},\, \alpha)} < \tau \right\}, \tag{2}$$

where D_{gt} is the ground-truth depth maps and D_{moge} is the aligned depth map predicted by MoGev2. We set $\alpha = 0.5$ to prevent instability from small depth values, and use $\tau = 0.1$ as the threshold. The resulting mask is then used to filter out low-confidence, noisy, or incomplete measurements from the ground-truth depth. We incorporate this into the training by adding a prior-guided depth term to the overall depth loss $\mathcal{L}_{\text{depth}}^{\text{vggt}}$ from VGGT [54]:

$$\begin{aligned} \mathcal{L}_{\text{depth}}^{p} &= \mathcal{L}_{\text{grad}}(\hat{D}^{M_{\text{conf}}}, D_{\text{moge}}^{M_{\text{conf}}}), \\ \mathcal{L}_{\text{depth}} &= \mathcal{L}_{\text{depth}}^{\text{vggt}} + \mathcal{L}_{\text{depth}}^{p} \end{aligned} \tag{3}$$

where $\hat{D}^{M_{\text{conf}}}$ denotes the predicted depth maps masked by the confidence prior, and $D_{\text{moge}}^{M_{\text{conf}}}$ corresponds to the filtered depth prediction from MoGe. $\mathcal{L}_{\text{grad}}$ is the gradient-based loss as implemented in VGGT. By restricting supervision only to high-confidence regions, our model can avoid overfitting to corrupted data, thereby achieving more accurate and stable depth estimation.

Dense Semantic Feature Fusion. While monocular or binocular models [13, 29, 57–59] can produce sharp and detailed geometry for a single view, multiview models tend to favor smoother predictions to preserve 3D consistency, which leads to the loss of fine geometric details. To address this, we fuse the globally aligned 3D feature f_{3d} from the backbone with dense semantic features f_s extracted from each input image using DINOv2 [38], providing additional local geometry cues that help produce sharper and more accurate predictions. These two features are concatenated along the feature dimension and passed through the DPT heads to regress the final depth and surface normals:

$$f_n = f_{3d} \oplus f_s, \tag{4}$$

where f_n is the input feature of the normal prediction head. We empirically validated this claim in the *ablation study*.

3.2. Mixture-of-Experts Design

Leveraging large vision models to predict multiple 3D geometric quantities has proven to be effective for several downstream tasks. However, a single decoder feature may be insufficient to capture the varying domain characteristics of different 3D scenarios. To address this limitation, we draw inspiration from the Mixture-of-Experts (MoE) framework, which has shown strong scalability and efficiency in large language models [9, 41, 65–67]. In the MoE design, multiple experts act as independent sub-networks trained to capture distinct aspects of the data. Building on this idea, we propose an MoE framework for 3D geometric vision as demonstrated in Fig. 2. This framework dynamically routes features to domain-specific experts, allowing the backbone to learn specialized representations and achieve substantial improvements across different domains. Subsequently, we formulate the MoE forward pass and training objectives to fully leverage the MoE framework in 3D visual geometry reconstruction.

MoE Forward-Pass. Within the MoE framework, an MoE layer serves as a modular component that enables conditional computation and expert specialization. A typical MoE layer comprises multiple feed-forward networks (FFNs), each serving as an expert. For initialization, we replicate the FFNs from the alternating attention structure (global and frame attention) to construct an ensemble of experts ε_i. A router is then employed to predict the assignment probability of each token to the corresponding experts. In our framework, the router is implemented as a linear layer, and the routing process can be expressed as:

$$\mathcal{P}(x)_i = \frac{e^{f(x)_i}}{\sum_j^E e^{f(x)_j}}, \tag{5}$$

where E represents the number of experts and $f(x) = W \cdot x$ is the weight logits produced by the router, and the W denotes the lightweight training parameters. Each token is then processed by the top-K experts with the highest assignment probabilities, and the final representation is obtained as the weighted sum according to these probabilities:

$$\text{MoE}(x) = \sum_{i=1}^{K} \mathcal{P}(x)_i \cdot \varepsilon(x)_i. \tag{6}$$

MoE Training Objectives. Since integrating multiple experts may lead to uneven expert utilization, it is necessary to apply load-balancing constraints to the MoE layer to regularize training. We therefore incorporate the differentiable load-balancing loss [14, 31] in each MoE layer to encourage all experts to process tokens in a balanced manner:

$$\mathcal{L}_{\text{moe}} = E \cdot \sum_{i=1}^{E} \mathcal{F}_i \cdot \mathcal{G}_i, \tag{7}$$

where $\mathcal{F}_i$ denotes the fraction of tokens processed by each expert ε_i, and $\mathcal{G}_i$ represents the average routing probability of ε_i, which can be formulated as:

$$\begin{aligned} \mathcal{F} &= \frac{1}{K}\sum_{i=1}^{E}\mathbf{1}\{\text{argmax}\mathcal{P}(x)=i\}, \\ \mathcal{G} &= \frac{1}{K}\sum_{i=1}^{K}\mathcal{P}(x)_i. \end{aligned} \tag{8}$$

3.3. Multi-task Training Objectives

Building on the training objectives of VGGT [54], we initially employ the pointmap loss $\mathcal{L}_{\text{points}}$, the camera loss $\mathcal{L}_{\text{cam}}$, and the tracking loss $\mathcal{L}_{\text{track}}$. We then introduce the depth loss $\mathcal{L}_{\text{depth}}$ after our confidence-based depth refinement and additional loss terms to further improve the accuracy and generalization of the model.

(1) Local Point Loss $\mathcal{L}_{\text{pts\_local}}$. Monocular geometry estimation often suffers from focal-distance ambiguity. To address this, we employ an additional local point loss to improve monocular depth estimation following [57, 62]. Given each image $\mathbf{I}_i$, a local point map $\hat{\mathbf{P}}_i$ is formed based on the predicted depth and focal parameters following VGGT [54]. During training, this predicted point map can be aligned with the ground truth point map by solving for a single optimal scale factor $\hat{s}$ that minimizes the depth-weighted $\mathcal{L}_1$ distance over the entire image sequence. The optimization problem is formulated as:

$$\hat{s} = \arg\min_{s}\sum_{i=1}^{N}\sum_{j=1}^{H\times W}\frac{1}{z_{i,j}}\|s\hat{\mathbf{p}}_{i,j}-\mathbf{p}_{i,j}\|_1, \tag{9}$$

where $\hat{\mathbf{p}}_{i,j}\in\mathbb{R}^3$ denotes the predicted 3D point at index j of the point map $\hat{\mathbf{P}}_i$, and $\mathbf{p}_{i,j}$ is the corresponding ground-truth in $\mathbf{P}_i$. The term $z_{i,j}$ corresponds to the ground-truth depth, which is the z-component of $\mathbf{x}_{i,j}$. Finally, the local point cloud reconstruction loss $\mathcal{L}_{\text{points}}$ is defined based on the optimal scale factor$\hat{s}$:

$$\mathcal{L}_{\text{pts\_local}} = \sum_{i=1}^{N}\sum_{j=1}^{H\times W}\frac{1}{z_{i,j}}\|\hat{s}\hat{\mathbf{x}}_{i,j}-\mathbf{x}_{i,j}\|_1. \tag{10}$$

(2) Point Normal Loss $\mathcal{L}_{\text{pts\_n}}$. To encourage the reconstruction of locally smooth surfaces, we employ the normal loss proposed by MoGe [57]. For each point $\hat{\mathbf{X}}_i$ in the predicted point map, its normal vector $\hat{\mathbf{n}}_{i,j}$ is computed from the cross product of the vectors to its neighboring points on the image grid. The predicted normals are then supervised by minimizing the angular difference with their corresponding ground-truth normals $\mathbf{n}_{i,j}$:

$$\mathcal{L}_{\text{pts\_n}} = \sum_{i=1}^{N}\sum_{j=1}^{H\times W}\arccos\left(\hat{\mathbf{n}}_{i,j}\cdot\mathbf{n}_{i,j}\right). \tag{11}$$

(3) Predicted Normal Loss $\mathcal{L}_{\text{n}}$. In addition to the pointmap normal loss, we introduce a predicted normal loss to supervise the normal head, which is responsible for predicting view-space normals:

$$\mathcal{L}_{\text{n}} = \mathcal{L}_1(N,\bar{N}), \tag{12}$$

where N is the ground-truth normal and $\bar{N}$ is the view-space normal produced by the normal prediction head.

3.4. Model Training

We train our model end-to-end using the following multi-task training objective:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{pts}} + \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \lambda_{\text{track}}\mathcal{L}_{\text{track}} \\ + \lambda_{\text{moe}}\mathcal{L}_{\text{moe}} + \lambda_{\text{pts\_local}}\mathcal{L}_{\text{pts\_local}} + \lambda_{\text{pts\_n}}\mathcal{L}_{\text{pts\_n}} + \lambda_{\text{n}}\mathcal{L}_{\text{n}} \end{aligned} \tag{13}$$

Our model is trained on a diverse dataset with varying levels of quality. Consequently, inaccurate annotations can occasionally cause unstable loss spikes during the training process. To mitigate this issue, we implement an adaptive loss strategy to **stabilize training**. Specifically, we maintain a sliding window of recent loss values and compute their mean μ and standard deviation σ. A dynamic threshold T_L is then defined following the k-sigma rule:

$$T_L = \mu_L + k\sigma_L, \tag{14}$$

where we set $k = 3$ by default in our experiments. If the current loss exceeds this threshold $L_{cur} > T_L$, it is considered an outlier and clipped to the threshold. This ensures that the training is guided by the typical distribution of losses rather than being dominated by rare extreme values. By continuously updating the loss history and applying this strategy, we effectively mitigate optimization instability while preserving the overall learning dynamics.

As for implementation details, we train our model based on the pretrained VGGT checkpoint and set the hyperparameter as follows: $\lambda_{\text{moe}} = 0.01, \lambda_{\text{pts\_local}} = 0.5, \lambda_{\text{pts\_n}} = 1.0, \lambda_{\text{n}} = 1.0$. We adopt the same training dataset as VGGT and extend it with an internal dataset that spans indoor, outdoor, object-centric, human-centric, and dynamic scenes.

4. Experiments

We compare our method to state-of-the-art approaches on a variety of 3D geometric prediction tasks to demonstrate its effectiveness and robustness.

4.1. Pointmap Estimation

For pointmap evaluation, we evaluate the model on the object-centric DTU [22] and scene-level ETH3D [46] datasets, sampling keyframes every 5 images. We further follow the evaluation protocols from [56, 62] and assess the quality of reconstructed multi-view point maps on

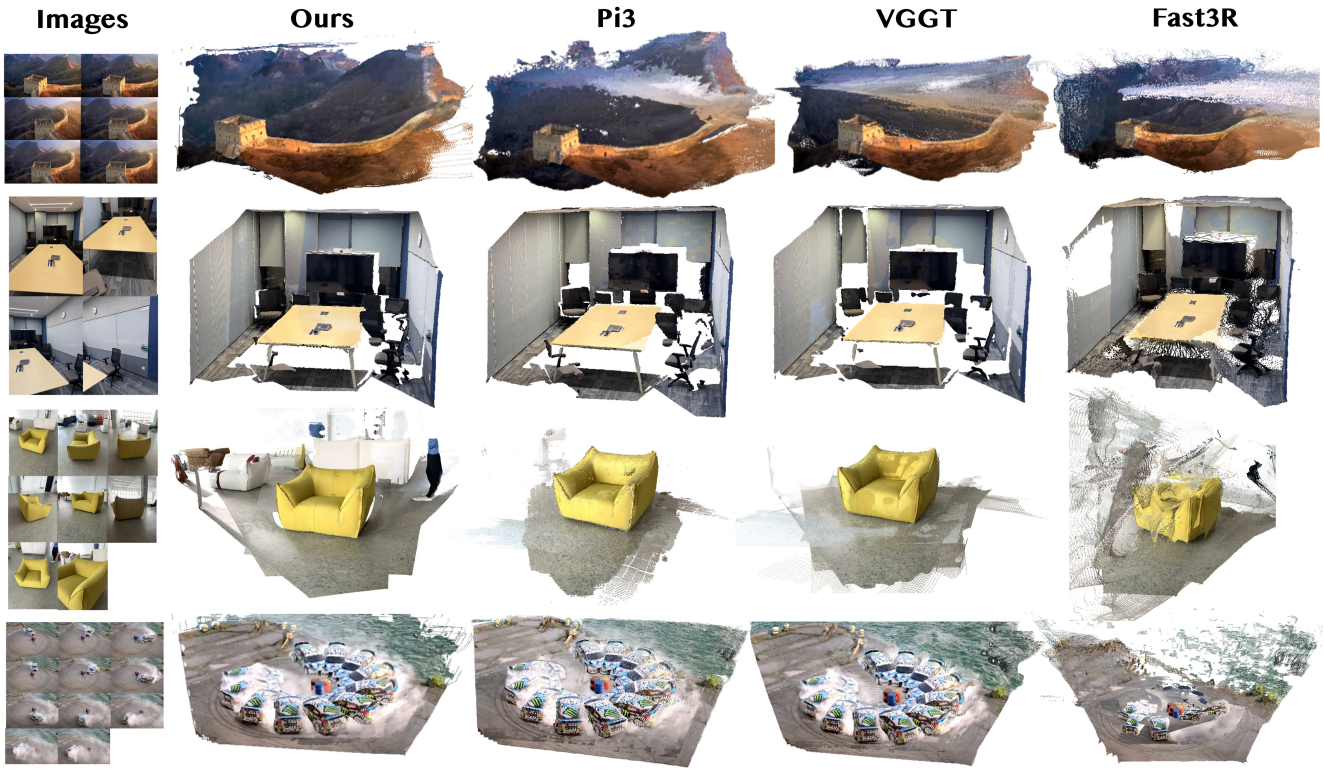

Figure 4. Qualitative comparison of multi-view 3D reconstruction. Our method demonstrates superior accuracy and robustness across diverse scenarios compared to previous feed-forward approaches.

the 7-Scenes [48] and NRGBD [2] datasets. Keyframes are selected with a stride of 200 for 7-Scenes and 500 for NRGBD. The predicted point maps are first aligned to the ground truth using the Umeyama algorithm for coarse Sim(3) alignment, and then refined using Iterative Closest Point (ICP). We report Accuracy (Acc.), Completion (Comp.), and Normal Consistency (N.C.) in Tables 1. We also present qualitative comparisons for pointmap estimation in Fig. 4. The results demonstrate the effectiveness of our method across various 3D reconstruction scenarios and show consistently high performance in both sparse-view and dense-view settings. It is noteworthy that Pi3 [62] often produces "checkerboard" artifacts due to insufficient learning in its transformer architecture. VGGT [54] and Fast3R [68] tend to yield less accurate reconstructions and struggle to generalize across scenarios. In contrast, our method reconstructs geometry that is both accurate and spatially consistent across diverse conditions.

4.2. Monocular Depth Estimation

For monocular depth estimation, we benchmark on Sintel [5], Bonn [39], and NYUv2 [49] and adopt Absolute Relative Error (Abs Rel ↓) and threshold accuracy ($\delta < 1.25$ ↑) as evaluation metrics [57, 62, 75]. As reported in Tab. 2, our method achieves high-quality results among multi-view approaches and it demonstrates comparable performance with monocular depth estimation model [57, 58].

4.3. Camera Pose Estimation

For camera pose estimation, we utilize both angular accuracy and distance error for evaluation.

Angular Accuracy Metrics. We evaluate predicted poses on the scene-level RealEstate10K [78] and object-centric Co3Dv2 [44] datasets following VGGT [54]. For each sequence, we randomly sample 10 images and form all possible pairs. We then compute angular errors of relative rotations and translations, reporting Relative Rotation Accuracy (RRA ↑) and Relative Translation Accuracy (RTA ↑). We also report a unified metric as the Area Under the Curve (AUC ↑) of the min(RRA, RTA)–threshold curve. Results in Tab. 3 demonstrate that our method sets a new state of the art on RealEstate10K in the zero-shot setting, while remaining competitive with the best-performing methods on the in-domain Co3Dv2 benchmark.

Distance Error Metrics. We further assess performance using Absolute Trajectory Error (ATE ↓), Relative Pose Error for translation (RPE trans ↓), and rotation (RPE rot ↓), following CUT3R [56] using the benchmarks TUM-Dynamics [51].. Predicted trajectories are first aligned with ground truth via a Sim(3) transformation before error com-

Method	DTU [22]						ETH3D [46]						7-Scenes [48]						NRGBD [2]					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Fast3R [68]	3.340	1.919	2.929	1.125	0.671	0.755	0.832	0.691	0.978	0.683	0.667	0.766	0.038	0.015	0.056	0.018	0.645	0.725	0.072	0.030	0.050	0.016	0.790	0.934
CUT3R [56]	4.742	2.600	3.400	1.316	0.679	0.764	0.617	0.525	0.747	0.579	0.754	0.848	0.022	0.010	0.027	0.009	0.668	0.762	0.086	0.037	0.048	0.017	0.800	0.953
FLARE [76]	2.541	1.468	3.174	1.420	0.684	0.774	0.464	0.338	0.664	0.395	0.744	0.864	0.018	0.007	0.027	0.014	0.681	0.781	0.023	0.011	0.018	0.008	0.882	0.986
VGGT [54]	1.338	0.779	1.896	0.992	0.676	0.766	0.280	0.185	0.305	0.182	0.853	0.950	0.022	0.008	0.026	0.013	0.663	0.757	0.017	0.010	0.015	0.005	0.893	0.988
Pi3 [62]	1.198	0.646	1.849	0.607	0.678	0.768	0.194	0.131	0.210	0.128	0.883	0.969	0.015	0.007	0.022	0.011	0.687	0.790	0.015	0.008	0.013	0.005	0.898	0.987
Ours	1.011	0.584	1.482	0.619	0.695	0.782	0.234	0.141	0.265	0.141	0.868	0.970	0.015	0.006	0.026	0.010	0.694	0.797	0.012	0.006	0.015	0.005	0.907	0.992

Table 1. **Point Map Estimation.** Keyframes are selected every 5 images for DTU and ETH3D, 40 images for 7-Scenes and 100 images for NRGBD. We present the accuracy (Acc.), completion (Comp.) and normal consistency (N.C.) as the evaluation metrics with each cell colored to indicate the best and the second.

Method	Sintel [5]		Bonn [39]		NYU-v2 [49]	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
DUSt3R [59]	0.488	0.532	0.139	0.832	0.081	0.909
MASt3R [29]	0.413	0.569	0.123	0.833	0.110	0.865
MonST3R [75]	0.402	0.525	0.069	0.954	0.094	0.887
Fast3R [68]	0.544	0.509	0.169	0.796	0.093	0.898
CUT3R [56]	0.418	0.520	0.058	0.967	0.081	0.914
FLARE [76]	0.606	0.402	0.130	0.836	0.089	0.898
VGGT [54]	0.335	0.599	0.053	0.970	0.056	0.951
MoGe [57]	0.273	0.695	0.050	0.976	0.055	0.952
Pi3 [62]	0.277	0.614	0.044	0.976	0.054	0.956
Ours	0.274	0.620	0.050	0.977	0.051	0.957

Table 2. **Monocular Depth Estimation on Sintel, Bonn, and NYU-v2.** We present the absolute relative error (Abs Rel) and threshold accuracy ($\delta < 1.25$) as the evaluation metrics with each cell colored to indicate the best and the second.

Method	RealEstate10K [78]			Co3Dv2 [44]			TUM-dynamics [51]		
	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	ATE ↓	RPE trans ↓	RPE rot ↓
Fast3R [68]	99.05	81.86	61.68	97.49	91.11	73.43	0.090	0.101	1.425
CUT3R [56]	99.82	95.10	81.47	96.19	92.69	75.82	0.047	0.015	0.451
FLARE [76]	99.69	95.23	80.01	96.38	93.76	73.99	0.026	0.013	0.475
VGGT [54]	99.97	93.13	77.62	98.96	97.13	88.59	0.012	0.010	0.311
Pi3 [62]	99.99	95.62	85.90	99.05	97.33	88.41	0.014	0.009	0.312
Ours	99.98	95.47	86.28	99.08	97.36	88.42	0.010	0.009	0.305

Table 3. **Camera Pose Estimation on RealEstate10K, Co3Dv2 and TUM-dynamics.** We present the metrics that measure the ratio of angular accuracy of rotation/translation under an error of 30 degrees for RealEstate10K and Co3Dv2. We also present the distance error of rotation/translation for TUM-dynamics with each cell colored to indicate the best and the second.

putation. As reported in Tab. 3, our method delivers more accurate results than previous approaches across both synthetic and real-world scenarios.

4.4. Normal Estimation

For normal estimation, we evaluate our model on several benchmark datasets spanning both indoor and outdoor scenes. We compare against Magrigold [25], Lotus [19], GeoWizard [16], and StableNormal [71]. As shown in Tab.4, our model achieves superior performance across multiple benchmarks. By alleviating the inherent ambiguity of monocular estimation and incorporating dense semantic features, our method produces more reliable and detailed predictions across diverse scenarios.

Method	NYUv2 [49]			ScanNet [73]			IBims-1 [27]		
	Mean ↓	Med ↓	$\delta_{11.25^\circ}$ ↑	Mean ↓	Med ↓	$\delta_{11.25^\circ}$ ↑	Mean ↓	Med ↓	$\delta_{11.25^\circ}$ ↑
Marigold [25]	20.8	11.1	50.4	21.2	12.2	45.6	18.4	8.4	64.7
Lotus [19]	17.5	8.6	58.7	18.1	8.8	58.2	19.2	5.6	66.2
GeoWizard [16]	20.4	11.9	47.0	21.4	13.9	37.1	19.7	9.7	58.4
StableNormal [71]	19.7	10.5	53.0	18.1	10.1	56.0	17.2	8.1	66.7
Ours	15.1	7.3	63.5	16.1	7.2	64.4	15.0	4.2	72.6

Table 4. **Normal Estimation on NYUv2, Scannet, IBims-1**. We report the mean and median angular errors with each cell colored to indicate the best and the second.

Model	DTU [22]			NYUv2 [49]		RealEstate10K [78]		
	Acc. ↓	Comp. ↓	N.C. ↑	Abs Rel ↓	$\delta < 1.25$ ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑
w/o $\mathcal{L}$, w/o MoE	1.338	1.896	0.676	0.056	0.951	99.97	93.13	77.62
w/o MoE	1.297	1.625	0.682	0.054	0.953	99.94	94.27	85.14
Ours	**1.011**	**1.482**	**0.695**	**0.051**	**0.957**	**99.98**	**95.47**	**86.28**

Table 5. **Ablation study** on the key components of our model. The results illustrate how performance metrics improve progressively as each component is incorporated into the baseline. Note that we report the *Mean* values for the ETH3D dataset.

4.5. Ablation Study

Mixtures-of-Experts. To validate the effectiveness of our proposed framework, we perform an ablation study by removing key components and training corresponding model variants. We compare the model variants by evaluating their performance on pointmap estimation (DTU [22]), monocular depth prediction (NYUv2 [49]), and camera pose (RealEstate10K [78]). As shown in Tab. 5, the results confirm that both the MoE and the tailored loss (including the confidence-based depth refinement) play a crucial role, contributing significantly to the overall performance of our model. It is noteworthy that all model variants, including the baseline, are trained for the same number of steps.

Confidence-based Depth Refinement. For depth prediction, we design a confidence-based depth refinement that utilizes MoGev2 to filter depth supervision. By restricting supervision to high-confidence regions, our model avoids overfitting to corrupted data and achieves more accurate depth estimation, which can be shown in Tab. 5 and Fig. 5.

Dense Semantic Feature Fusion. MoRE proposes to fuse globally aligned 3D backbone features with dense semantic features for normal predictions. This feature fusion allows the model to capture both local geometric details and global

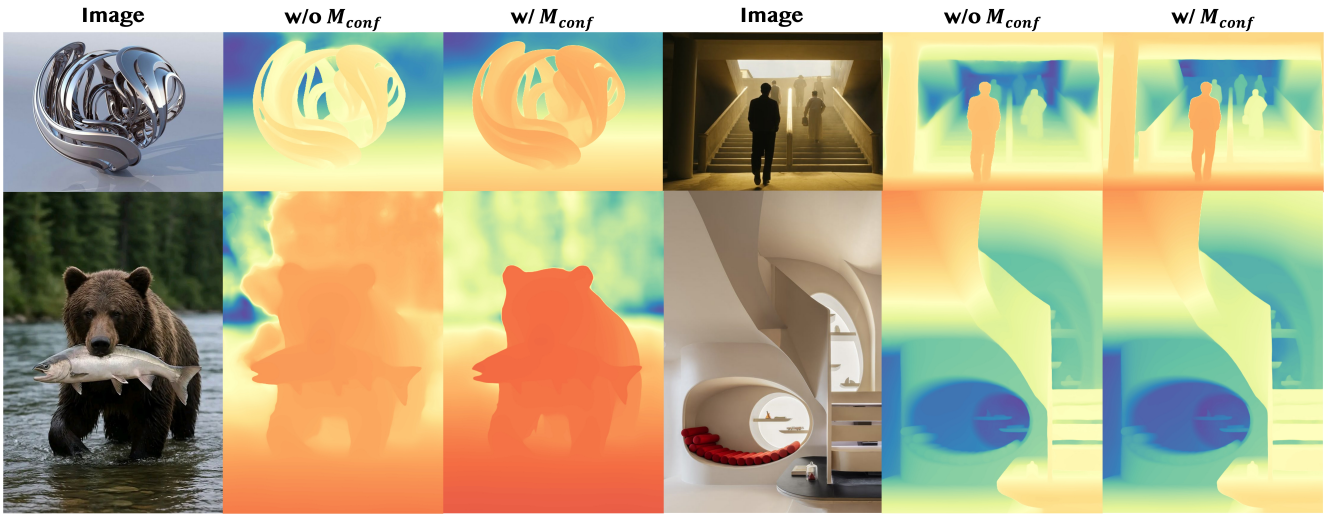

Figure 5. Ablation for confidence-based depth refinement. We demonstrate the effectiveness of the confidence-based depth refinement for more accurate depth estimation.

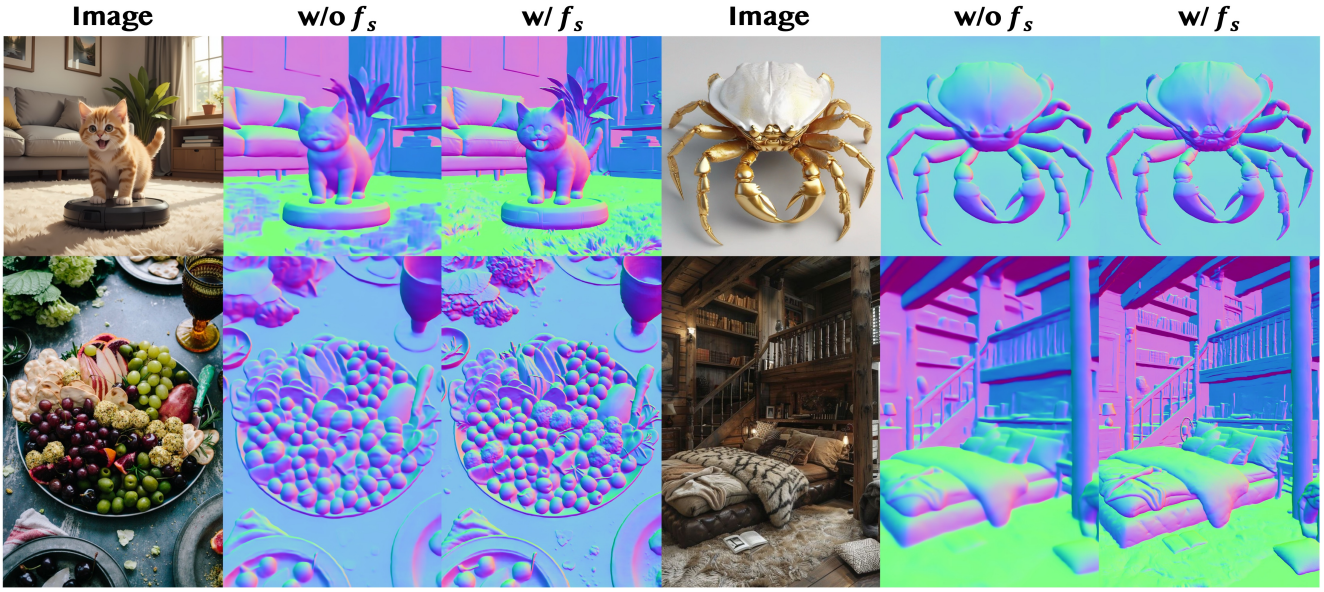

Figure 6. Ablation for dense semantic feature fusion. We demonstrate the effectiveness of the dense semantic feature fusion for sharper and more accurate normal estimation.

contextual cues, leading to sharper and more reliable predictions. As shown in Fig. 6, it further strengthens the representation of fine-grained structures, which is essential for high-fidelity 3D reconstruction.

5. Conclusion

We introduce MoRE, a large-scale 3D foundation model built upon a Mixture-of-Experts (MoE) framework for 3D visual geometry reconstruction. Unlike previous methods that rely on a single shared representation for all scenarios, MoRE dynamically routes features to domain-specific experts, thereby enhancing representational capacity and improving prediction accuracy. To address the noise and inconsistency of real-world data, our model employs a confidence-based depth refinement module, which effectively enhances the reliability of depth estimation. We further introduce a dense semantic fusion that fuses local geometry features with 3D backbone features for more detailed normal estimation. The model is trained with tailored multi-task losses and an adaptive loss mechanism to stabilize training across diverse datasets. Extensive experiments demonstrate that MoRE achieves state-of-the-art performance across multiple benchmarks, providing a versatile and scalable backbone for downstream applications.

Acknowledgment

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 25511103000, and NSFC (62571314).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Dejan Azinovic, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, pages 6280–6291, 2022. 6, 7

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5835–5844, 2021. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, pages 19697–19705, 2023. 2

[5] Aljaz Bozic, Pablo R. Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular RGB scene reconstruction using transformers. In *NeurIPS*, pages 1403–1414, 2021. 6, 7

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, 2020. 2

[7] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *CVPR*, pages 1050–1060, 2025. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[9] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *ACL*, pages 1280–1297, 2024. 2, 3, 4

[10] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences. In *ICCV*, 2025. 2

[11] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, pages 5547–5569, 2022. 2

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2

[13] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, and Chengfei Lyu. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025. 2, 3, 4

[14] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022. 4

[15] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 2

[16] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, pages 241–258, 2024. 7

[17] Jingnan Gao, Zhuo Chen, Xiaokang Yang, and Yichao Yan. Anisdf: Fused-granularity neural surfaces with anisotropic encoding for high-fidelity 3d reconstruction. In *ICLR*, 2025. 2

[18] 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*, pages 2495–2504, 2020. 2

[19] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *ICLR*, 2025. 7

[20] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models (2022). *arXiv preprint arXiv:2203.15556*, 2022. 2

[21] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *CVPR*, pages 1071–1081, 2025. 2

[22] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, pages 406–413, 2014. 5, 7

[23] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 3

[24] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang

Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *ICLR*, 2025. 3

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 7

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 2

[27] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. *arXiv preprint arXiv:1805.01328*, 2018. 7

[28] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In *ICLR*, 2021. 2

[29] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, pages 71–91, 2024. 2, 4, 7

[30] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Unimoe: Scaling unified multimodal llms with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439, 2025. 3

[31] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 3, 4

[32] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. *CVPR*, 2024. 2

[33] Yueen Ma, Yuzheng Zhuang, Jianye Hao, and Irwin King. 3d-moe: A mixture-of-experts multi-modal LLM for 3d vision and pose diffusion via rectified flow. *arXiv preprint arXiv:2501.16698*, 2025. 3

[34] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021. 2

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[36] Takeshi Noda, Chao Chen, Weiqi Zhang, Xinhai Liu, Yu-Shen Liu, and Zhizhong Han. Multipull: Detailing signed distance functions by pulling multi-level queries at multi-step. *NeurIPS*, 37:13404–13429, 2024.

[37] Takeshi Noda, Chao Chen, Junsheng Zhou, Weiqi Zhang, Yu-Shen Liu, and Zhizhong Han. Learning bijective surface parameterization for inferring signed distance functions from sparse point clouds with grid deformation. In *CVPR*, pages 22139–22149, 2025. 2

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 2, 4

[39] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IROS*, pages 7855–7862, 2019. 6, 7

[40] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *CVPR*, pages 8645–8654, 2022. 2

[41] Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models. In *ACL*, pages 5005–5018, 2025. 2, 4

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[43] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale. In *ICML*, pages 18332–18346, 2022. 3

[44] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, pages 10881–10891, 2021. 6, 7

[45] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 2

[46] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 2538–2547, 2017. 5, 7

[47] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Y. Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *arXiv preprint arXiv:2305.14705*, 2023. 3

[48] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, pages 2930–2937, 2013. 6, 7

[49] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob

Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760, 2012. 6, 7

[50] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[51] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, pages 573–580, 2012. 6, 7

[52] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, pages 5283–5293, 2025. 2

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[54] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7

[55] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, pages 27171–27183, 2021. 2

[56] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2, 5, 6, 7

[57] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, 2025. 2, 4, 5, 6, 7

[58] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 3, 4, 6

[59] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024. 2, 4, 7

[60] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023. 3

[61] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *ICCV*, 2023. 2

[62] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 2, 3, 5, 6, 7

[63] Zipeng Wang and Dan Xu. Flashvggt: Efficient and scalable visual geometry transformers with compressed descriptor attention. *arXiv preprint arXiv:2512.01540*, 2025. 2

[64] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. In *ICML*, 2024. 3

[65] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2, 3, 4

[66] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[67] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 3, 4

[68] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *CVPR*, 2025. 2, 6, 7

[69] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018. 2

[70] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, pages 4805–4815, 2021. 2

[71] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Trans. Graph.*, 43(6): 250:1–250:18, 2024. 7

[72] Gokul Yenduri, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Rutvij H Jhaveri, Weizheng Wang, Athanasios V Vasilakos, Thippa Reddy Gadekallu, et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*, 2023. 2

[73] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, pages 12–22, 2023. 7

[74] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *CVPR*, 2024. 2

[75] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025. 2, 6, 7

[76] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *CVPR*, 2025. 2, 7

[77] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *CVPR*, pages 21508–21518, 2023. 2

[78] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 65, 2018. 6, 7

[79] Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. Llama-moe: Building mixture-of-experts from llama with continual pre-training. In *EMNLP*, pages 15913–15923, 2024. 3