VGGT-Ω

Jianyuan Wang[1,2] Minghao Chen[1] Shangzhan Zhang[1] Nikita Karaev[1]
Johannes Schönberger[2] Patrick Labatut[2] Piotr Bojanowski[2] David Novotny
Andrea Vedaldi[1,2] Christian Rupprecht[1]

[1]Visual Geometry Group, University of Oxford [2]Meta AI

Abstract

We introduce VGGT-Ω, a feed-forward model for 3D reconstruction that improves accuracy, efficiency, and capabilities for both static and dynamic scenes. Prior models such as VGGT have shown that feed-forward 3D reconstruction is, in many cases, competitive with traditional optimization-based methods. Here, we show that the accuracy and robustness of these models scale predictably with model capacity and data size. To enable training 3D reconstruction models at an unprecedented scale, we introduce architectural changes that improve training efficiency and scalability, a high-quality data annotation pipeline that supports dynamic scenes, and a self-supervised learning protocol. We significantly simplify VGGT's architecture by using a single dense prediction head with multi-task supervision, removing expensive high-resolution convolutional layers, and introducing efficient scene tokens for feature aggregation in lieu of global attention. These changes allow us to train VGGT-Ω with 15× more supervised data than prior work and to leverage vast amounts of unlabeled videos, while requiring only ∼30% of VGGT's training memory. VGGT-Ω achieves strong results for 3D reconstruction of static and dynamic scenes across multiple benchmarks, e.g.*, improving over the previous best camera estimation accuracy by 77% on Sintel.*

1. Introduction

Recent work [51, 115, 119, 121] has demonstrated that feed-forward 3D reconstruction models can, in many cases, match and even surpass traditional structure-from-motion (SfM) pipelines [34, 87, 94]. However, while SfM has been developed for decades, feed-forward reconstruction models are new and relatively unexplored. In particular, the power of scale in machine learning is now well understood in other domains [45, 93, 137], but less so for 3D/4D vision. In this paper, we thus ask *whether we can scale up feed-forward reconstruction models, and what the benefits are.* To answer this question, we introduce VGGT-Ω, a model that scales 3D machine learning to significantly larger data and, optionally, model sizes than prior work.

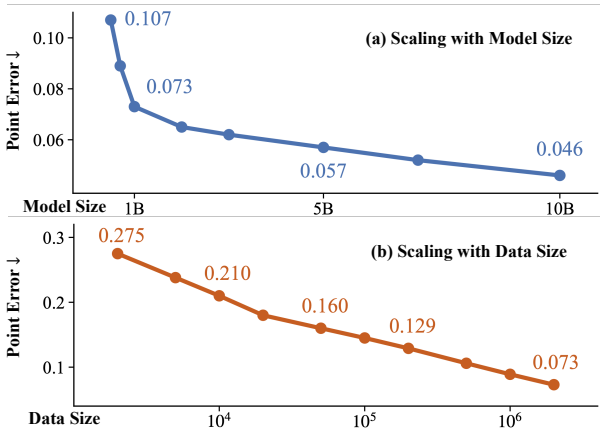

Figure 1. **Performance Gains from Data and Model Parameter Scaling**. As model size increases from 0.5B to 10B parameters and data scale grows from 2K to 2M sequences, performance improves consistently (measured by 3D point error, lower is better; note the different scaling of the axes). All models are trained on approximately the same number of tokens and evaluated by the average of six datasets (see Sec. 4.2 for details).

Scaling 3D vision introduces substantial technical challenges. We address these by: (i) revisiting and simplifying VGGT's architecture and training protocol to improve efficiency and enable large-scale training, (ii) introducing a scalable and highly accurate 3D annotation pipeline that also extends to dynamic content, and (iii) proposing a self-supervised training protocol to further increase the amount of training data available to the model.

We start by revisiting VGGT's architecture and training protocol to improve efficiency and scale models and data. First, we note that global attention is the main computational bottleneck in VGGT but its attention maps are very

sparse, consistent with recent findings [90, 110]. We take advantage of this fact by introducing a *scene-token attention* mechanism that involves only a small, fixed set of register tokens per frame while preserving scene-level context. Replacing 25% of global attention layers with scene-token attention incurs no measurable performance drop, whereas replacing all of them reduces the backbone's FLOPs to 6% of the original, though at the cost of a considerable performance drop. Second, we found that the high-resolution convolutional layers in dense prediction heads (*e.g.*, DPT) use a disproportionate amount of GPU memory for storing activations, despite accounting for only a small fraction of the model's parameters. Common techniques like FSDP and gradient checkpointing cannot mitigate this bottleneck. Instead, we show that replacing such layers with MLPs followed by pixel shuffle operators uses very little memory and maintains performance in benchmarks. Third, the original VGGT demonstrated that multi-task training (depth maps, point maps, and tracking features) is beneficial, but we find that additional dense *heads* are not necessary to enjoy these benefits. Instead, we only need to use corresponding multi-task *losses* based on a single dense head (depth) and sparse head (camera) for prediction. These three changes reduce memory usage to around 30% of the original model and modestly improve training and inference speed.

We also find that the diversity and quality of training data are critical. In particular, we argue that handling *dynamic* content is essential for scaling, as it unlocks orders of magnitude more Internet-like videos for training. While recent works [56, 116] leverage MegaSaM [57] to annotate dynamic videos, we empirically observe that, for the majority of videos, the resulting labels are not sufficiently accurate to serve as pseudo ground truth. Therefore, we develop a high-quality data annotation pipeline that can produce annotations for both rigid and dynamic videos at scale. The pipeline integrates VLM-based pre-filtering, VGGT, COLMAP, modern image-matching models, and supervised geometric post-filtering. Applied to around 40 million internal Internet-style videos, the filtering pipeline retains 0.8 million sequences with accurate annotations, roughly one-third of which contain dynamic content. Combined with existing datasets (both real and synthetic), this gives us in total 4M diverse scenes/sequences with precise 3D annotations, which is more than $15\times$ that of VGGT.

To further improve generalization, we introduce a self-supervised learning protocol inspired by DINO [74, 93]. We maintain teacher and student models initialized from a supervised VGGT-Ω checkpoint. Both models process the same input sequences under different augmentations and frame permutations. The student is trained to match the teacher's predictions and feature distribution (after aligning the frame order), while the teacher is updated via an exponential moving average of the student. We use this protocol to train VGGT-Ω on 18M unlabeled videos.

These improvements allow us to investigate the scaling properties of feed-forward reconstruction models. As illustrated in Fig. 1, we observe a consistent power-law-like improvement in reconstruction accuracy (measured by point error) as we increase the model capacity from 0.5B to 10B parameters and expand the training data from a few thousand to over two million different sequences.

VGGT-Ω delivers improved feed-forward reconstruction performance, achieving strong results across three static and three dynamic benchmarks. In particular, it compares favorably to post-optimization methods such as MegaSaM on dynamic scenes and to recent feed-forward methods such as Depth Anything 3 [59]. On Sintel, VGGT-Ω attains AUC@3° of 40.0 vs. 22.5 (by 77%) and AUC@30° of 79.1 vs. 58.3 (by 35%) for camera estimation, as well as $\delta_{1.25}$ of 93.5 vs. 74.1 (by 26%) for depth estimation, while being $50\times$ faster than MegaSaM. Furthermore, we show the potential of using reconstruction as a pre-training task and the effectiveness of our scene tokens in tasks like vision-language-action (VLA). To summarize, our contributions are:

- *Architecture:* we introduce scene-token attention, remove high-resolution convolutional layers, and eliminate redundant prediction heads (while maintaining multi-loss supervision) to enhance training efficiency and stability.
- *Scaling:* we explore the scaling behavior of feed-forward reconstruction models and show that performance improves consistently with both model size and data scale.
- *Dynamic reconstruction:* we reconstruct dynamic 3D content that contains complex non-rigid motion to support new applications and to learn from many more videos.
- *Data:* we introduce a new large-scale annotation and filtering pipeline to obtain high-quality annotations of static and dynamic 3D content. We also introduce a self-supervised training protocol to further scale data.

2. Related work

Static 3D Reconstruction. There is a long and rich history of research on 3D reconstruction, beginning with seminal works that established the theory of multi-view geometry [25, 35, 73, 75]. Follow-up work led to major practical advances, including robust SfM systems such as COLMAP [88] and other pipelines [2, 28, 65, 94]. In this paper, we focus on *feed-forward reconstruction models*, *i.e.*, neural networks that infer scene geometry and camera poses directly from one or more images. While recent SfM pipelines increasingly include learnable components such as keypoint detectors [21, 23, 108, 134] and feature matchers [14, 60, 85, 91], our work is most closely related to end-to-end differentiable SfM frameworks that learn geometry estimation directly [8, 102, 104, 105, 109, 112–

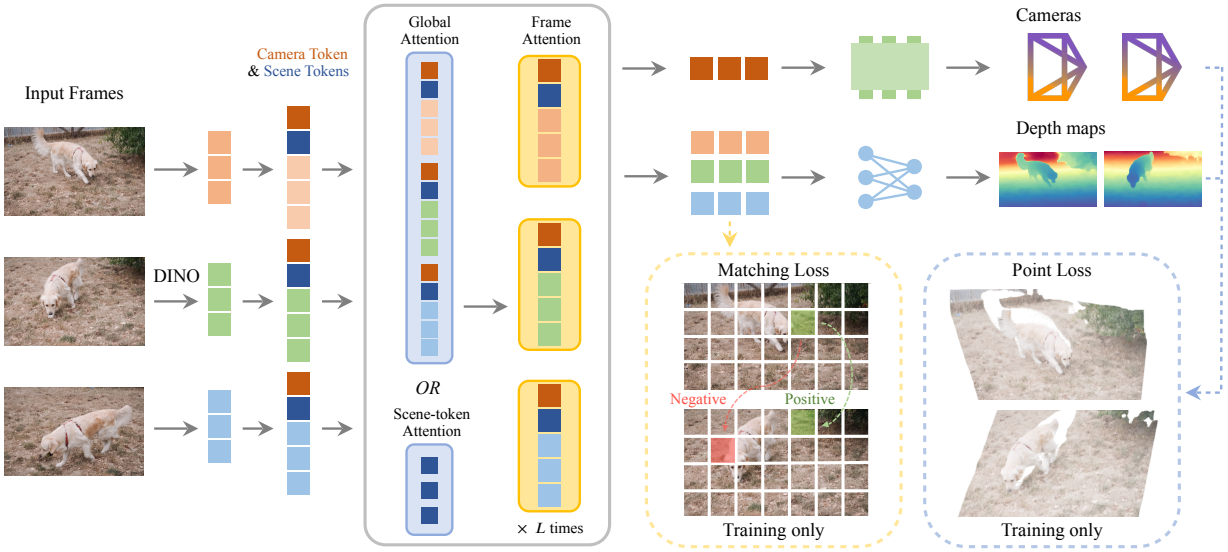

Figure 2. **Architecture Overview.** VGGT-Ω appends camera and scene tokens to the image tokens, and then alternates between global (or scene-token global) attention and frame attention layers. We replaced the redundant dense heads of VGGT with training-only losses.

114, 122, 144]. While these works demonstrate that end-to-end learning is possible in SfM, they still *combine* elements of classical SfM pipelines. DUSt3R [119] and its follow-up MASt3R [22] changed this by introducing fully feed-forward 3D reconstruction models that can estimate both scene geometry and camera parameters (extrinsics and intrinsics) directly from images. However, similar to multi-view stereo (MVS) approaches [29, 32, 68, 72, 79, 123, 131, 132, 140], the neural networks in DUSt3R and MASt3R operate only on image pairs and still require post-optimization to process additional views. A key improvement came from methods that process multiple images jointly, removing the need for optimization across views, including [24, 103, 111, 118, 124, 129, 139]. Among these, VGGT [115] is a representative approach that first surpassed post-optimization methods (*e.g.*, using bundle adjustment) while relying solely on feed-forward inference, prompting many follow-up works [13, 16, 19, 20, 38, 42, 44, 47, 49, 54, 58, 63, 69, 70, 81, 117, 121, 128, 135, 145].

Dynamic 3D Reconstruction. Monocular dynamic 3D reconstruction, or 4D reconstruction, aims to recover scene geometry that changes over time. This line of research also has a long history, with early work by Bregler et al. [9] and Torresani et al. [106]. Among recent contributions, MegaSaM [57] is particularly influential, combining feed-forward depth prediction with optimization-based non-rigid reconstruction, and ViPE [39] further builds on this direction. Several works have explored feed-forward 4D reconstruction with reduced optimization. MonST3R [138] and D^{2}USt3R [33] extend DUSt3R to handle pairs of images with dynamic 3D content. Align3R [67] builds on DUSt3R to infer cameras and align monocular depth predictions over time, though still relies on optimization beyond two views. CUT3R [118] and Point3R [124] support incremental reconstruction alongside dynamic scenes. Geo4D [43] takes a different approach, fine-tuning a video generator to recover 4D geometry. PI3 [121] adopts a VGGT-style model and trains it with dynamic data, while PAGE-4D [143] adapts VGGT through a module that separates static and moving regions. Human3R focuses on human-scene reconstruction [17]. SpatialTracker [126, 127], St4RTrack [26], DPM [98], and Uni4D [130] and others unify dynamic reconstruction and correspondence estimation.

3. Scaling Up Visual Geometry

With VGGT-Ω, our goal is to scale up feed-forward reconstruction models in terms of model capacity and training data, and to investigate the impact of these factors on reconstruction quality and generality. We begin by detailing VGGT-Ω's improvements to the architecture (Sec. 3.1) and training pipeline (Sec. 3.2) of VGGT. Then, we introduce a self-supervised training protocol that enables us to leverage unlabeled data (Sec. 3.4) and a new data pipeline to add high-quality annotations to millions of videos (Sec. 3.5).

3.1. A New Scalable Architecture

VGGT-Ω, illustrated in Fig. 2, is a feed-forward transformer f that maps N input images $I_1, \dots, I_N \in \mathbb{R}^{3\times H\times W}$ to corresponding cameras and depth maps:

$$((\boldsymbol{g}_1, D_1), \dots, (\boldsymbol{g}_N, D_N)) = f(I_1, \dots, I_N),$$

where $D_i \in \mathbb{R}^{H\times W}$ is the depth and $\boldsymbol{g}_i = (\boldsymbol{q}_i, \boldsymbol{t}_i, \boldsymbol{f}_i) \in \mathbb{R}^9$ is the concatenation of the rotation quaternion $\boldsymbol{q}_i \in \mathbb{R}^4$, the translation vector $\boldsymbol{t}_i \in \mathbb{R}^3$, and the field of view $\boldsymbol{f}_i \in \mathbb{R}^2$.

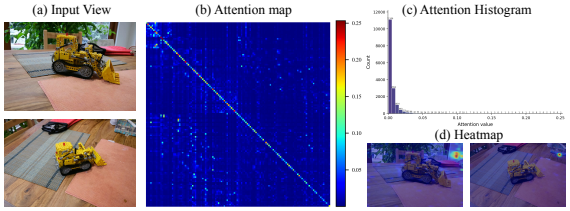

Figure 3. **Visualization of Global Attention in VGGT**. As shown in (b) global attention matrix, (c) its value distribution, and (d) spatial heatmaps, the attention (layer 13) is quite sparse.

As commonly done [87, 114, 115], we assume that the principal point is at the image center. The problem formulation is thus similar to VGGT [115], but it already departs from it, as the model *does not* predict point maps or tracking features explicitly (although it still supervises them, as discussed in Sec. 3.2). The network f encodes each image into tokens (Sec. 3.1.1), aggregates features across views with alternating-attention (Sec. 3.1.2), and maps the tokens to the final predictions with lightweight heads (Sec. 3.1.3).

3.1.1. Feature Extraction and Tokenization

We tokenize each image I_i with a DINOv3-initialized vision transformer [93], obtaining $\boldsymbol{z}_i^F = \mathrm{DINO}(I_i) \in \mathbb{R}^{H'W' \times C}$, where $H' = H/r$ and $W' = W/r$ for patch size r. For each image I_i, we also append one *camera token* $\boldsymbol{z}_i^{\text{cam}} \in \mathbb{R}^{1 \times C}$ and sixteen *scene tokens* $\boldsymbol{z}_i^{\text{scene}} \in \mathbb{R}^{16 \times C}$. The camera token is used to predict the camera parameters, and the scene tokens aggregate information about the scene. As in [115], these tokens can take one of two learnable parameters, one if image I_i is the *reference image* and the other otherwise. These are then concatenated to form the set of tokens $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_N) \in \mathbb{R}^{N \times (H'W'+17) \times C}$ where $\boldsymbol{z}_i = (\boldsymbol{z}_i^F, \boldsymbol{z}_i^{\text{cam}}, \boldsymbol{z}_i^{\text{scene}})$ are the tokens of image I_i.

3.1.2. Efficient Global Attention using Scene Tokens

Recall that in VGGT the tokens are processed by several layers of self-attention, which are, by definition, permutation equivariant. None of the tokens has an explicit encoding of the corresponding image identity (except for indicating whether a frame is the reference). To make the network aware of the different images, VGGT uses *alternating-attention* [115], alternating frame-wise attention within each image and global attention across all images. *Global* attention is the standard attention layer applied to all tokens $\boldsymbol{z}$, which we denote $\boldsymbol{z}' = \mathrm{attn}(\boldsymbol{z})$. *Frame-wise* attention is similar, but applied independently to $\boldsymbol{z}_i$ for each image I_i, which we denote $\boldsymbol{z}' = \mathrm{attn}_f(\boldsymbol{z}) = (\mathrm{attn}(\boldsymbol{z}_1), \dots, \mathrm{attn}(\boldsymbol{z}_N))$.

While effective, the global attention layers are computationally expensive, as they attend all tokens from all frames, with cost quadratic in the total number of tokens. However, we observe that global attention maps are typically sparse, as shown in Fig. 3, suggesting that global information might be exchanged efficiently through small bottlenecks. This is consistent with recent findings [90, 110]. We therefore replace 25% of the global attention layers with *scene-token attention* where global attention is restricted to the scene tokens only. Formally, we define $\boldsymbol{z}' = \mathrm{attn}_{\text{scene}}(\boldsymbol{z})$ where $(\boldsymbol{z}_1^{\text{scene}\prime}, \dots, \boldsymbol{z}_N^{\text{scene}\prime}) = \mathrm{attn}(\boldsymbol{z}_1^{\text{scene}}, \dots, \boldsymbol{z}_N^{\text{scene}})$, *i.e.*, only the scene tokens participate in self-attention. The information gathered by scene tokens is broadcast back to image tokens by subsequent attention layers. The standard global attention is interleaved with scene-token attention.

3.1.3. Decoding

The final set of tokens $\boldsymbol{z}' = (\boldsymbol{z}'_1, \dots, \boldsymbol{z}'_N)$ produced by the attention layers is decoded into depth maps and cameras.

Depth. In VGGT, all dense decoders are implemented with Dense Prediction Transformer (DPT) layers [82]. However, the final convolutional blocks in these DPT heads maintain several high-resolution feature maps, which are memory-intensive. To reduce this cost, we replace the blocks operating above $1/4$ of the input resolution with a lightweight upsampling head via a single MLP followed by a pixel-shuffle operator. The MLP outputs $2u^2$ channels ($u = 4$ in our implementation), and the pixel-shuffle operator rearranges them from $(H' \times W', 2u^2)$ to $(uH') \times (uW') \times 2$, where the two output channels correspond to depth and confidence.

Camera. The cameras $(\boldsymbol{g}_1, \dots, \boldsymbol{g}_N)$ are predicted jointly using a lightweight transformer over the camera tokens $(\boldsymbol{z}_1^{\text{Cam}}, \dots, \boldsymbol{z}_N^{\text{Cam}})$, followed by a final MLP. Unlike VGGT, our camera head predicts camera parameters in a single pass, without iterative refinement.

3.2. Training Losses

In VGGT, we found it beneficial to predict redundant dense heads (*e.g.*, point maps and tracks), but doing so is expensive during training. Instead, VGGT-Ω contains a single dense head for depth prediction. Although the model does not *directly* predict point maps and tracks, we *still* supervise these quantities via corresponding losses. We found that doing so results in nearly the same performance as using multiple dense prediction heads, while saving a lot of memory. To optimize the model, we thus use the loss:

$$\mathcal{L} = \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{point}}\mathcal{L}_{\text{point}} + \lambda_{\text{match}}\mathcal{L}_{\text{match}} \quad (1)$$

where λ_{cam}, λ_{depth}, λ_{point}, and λ_{match} are weights.

Camera loss. The camera loss $\mathcal{L}_{\text{cam}} = \sum_{i=1}^{N} |\hat{\boldsymbol{g}}_i - \boldsymbol{g}_i|$ compares the predicted cameras $\hat{\boldsymbol{g}}_i$ to the ground truth $\boldsymbol{g}_i$ using an ℓ_1 objective, which we found more stable than the Huber loss used in VGGT.

Depth loss. Following VGGT, the depth loss is formulated with aleatoric uncertainty and a gradient consistency term. Additionally, we account for the relative scale with respect to the ground truth. Therefore, we have: $\mathcal{L}_{\text{depth}} =$

Figure 4. **Qualitative Results.** VGGT-Ω handles both static and dynamic content, as evidenced by the overlaid traffic flow and the tennis player's trajectory. It also generalizes to hard scenes, *e.g.*, underwater coral reefs. Each example uses 64, 4, 9, 16, and 32 input frames.

$\sum_{i=1}^{N}\left[\left\|c_i^D \odot \left(1 + D_i^{-1}\right) \odot e_i\right\| + \left\|c_i^D \odot \nabla e_i\right\|\right] - \alpha \sum_{i=1}^{N} \log c_i^D$, where $e_i = \hat{D}_i - D_i$, c_i^D is the predicted uncertainty map, and $\odot$ denotes element-wise product.

Point loss. Point maps assign to each pixel the coordinates of the corresponding 3D point in the frame of the reference camera. The point maps can thus be inferred from the depth maps and the camera parameters via unprojection. Accordingly, our point loss $\mathcal{L}_{\text{point}}$ is the same as the depth loss $\mathcal{L}_{\text{depth}}$ up to replacing the residuals with $e_i = \pi^{-1}(\hat{D}_i, \hat{\boldsymbol{g}}_i) - P_i$, where π^{-1} denotes unprojection and P_i is point map.

Matching loss. The matching loss $\mathcal{L}_{\text{match}}$ is applied to the tokens output by the last attention layer. It pulls together features of positive token pairs (corresponding to the same 3D location) and pushes apart negative pairs: $\mathcal{L}_{\text{match}} = \mathbb{E}_{\text{pos}}\big[-\log \sigma(s)\big] + \mathbb{E}_{\text{neg}}\big[-\log(1-\sigma(s))\big]$, where s is the cosine similarity between ℓ_2-normalized tokens, σ is the sigmoid function, and $\mathbb{E}$ denotes averaging over positive and negative pairs, *i.e.*, a weighted binary cross-entropy. Details of how to construct the pairs are provided in the supplement.

3.3. Dynamic Reconstruction

By predicting depth maps and camera parameters, our architecture and training objectives naturally support reconstruction of dynamic scenes. This is important for applications and unlocks orders of magnitude more training data, as most videos contain scene motion. Some recent works [119] use point maps to represent and recover the camera parameters instead. This works in static scenes, but otherwise requires segmenting out dynamic pixels, as in MonST3R [138], or using extensions like dynamic point maps [99, 100].

3.4. Self-supervised Training

Inspired by DINO [74, 93], we use a teacher-student strategy for self-supervised learning with unlabeled videos. Specifically, we maintain a *student* network that is updated by gradient descent, and a *teacher* network that is updated only via an exponential moving average of the student network. Both networks are initialized from the VGGT-Ω checkpoint trained on supervised data. Given a video sequence, we feed the same set of frames to both networks,

but apply independent stochastic augmentations, including color jittering and blurring, random 90° rotations, random patch masking, and random frame-order permutation (which affects the selection of the reference frame). After restoring both streams to a common order, we require the student to match the teacher by: (a) An ℓ_2 feature-matching loss aligns the student's tokens with the teacher's across multiple layers. (b) Regression losses supervise camera and depth. To prevent collapse, the camera and depth heads are *frozen* during self-supervision. The teacher's parameters are updated as $\theta^{\mathrm{T}} \leftarrow m\,\theta^{\mathrm{T}} + (1-m)\,\theta^{\mathrm{S}}$ instead of gradient descent. This distillation scheme thus enforces invariance to appearance changes and frame order, enabling effective learning from millions of unlabeled videos.

3.5. Training Data

3.5.1. Data Sources

We first collect publicly available datasets: Aria series [76, 77], Bedlam [7], BEHAVIOR-1K [52], Co3Dv2 [83], uCo3D [66], DL3DV [61], Dynamic Replica [46], EDEN [50], EFM3D [96], HOT3D [6], Habitat [86], Hypersim [84], Mapfree [4], Mapillary Metropolis [71], MPSD [3], Megadepth [55], Megasynth [41], Mid-Air [27], Mvssynth [40], ParallelDomain-4D [36], Replica [95], SAIL-VOS [37], ScanNet Series [18, 133], TartanAirV2 [120], TartanGround [78], Taskonomy [136], UnrealStereo4K [107], Virtual KITTI [12], Waymo [101], and Wildrgbd [125]. We exclude Kubric [31] and PointOdyssey [142] used by VGGT because their background geometry is fake and yields invalid depth. Additionally, we use several internal datasets, which include artist-created object assets, rigid and dynamic synthetic environments, real-world device captures, and so on. For non-synthetic datasets (*e.g.*, those annotated by SfM pipelines), we remove noisy depth values via a multi-view consistency check and discard sequences with too few valid depth pixels. In total, these datasets contain approximately 3M sequences, each containing between 10 and 20,000 images.

3.5.2. Data Annotation Pipeline

To further expand our training data, we mine a large internal video collection of roughly 40M Internet-style videos. We first assess each video for suitability for 3D reconstruction, *e.g.* filtering out clips with large watermarks or abrupt shot changes. Videos that pass this check are used for self-supervised training as described in Sec. 3.4. We then introduce a new pipeline to annotate videos with camera parameters and depth maps, for both static and dynamic scenes. While we aim at obtaining a good yield, we prioritize annotation quality and reject data that is suspected to be low quality. This may exclude extreme camera motions or highly dynamic scenes, but these are still well represented in the synthetic datasets. Overall, we obtain a collection of about 200K dynamic scenes and 600K static scenes with high-quality camera and depth annotations.

VLM pre-filtering. We prompt a Vision-Language Model (VLM) to discard videos if they are unlikely to be reconstructed with multi-view geometry. The VLM classifies 50% of the clips as too difficult to reconstruct, *e.g.* containing multiple clips, extreme motion blur, or heavy overlays/watermarks. It also classifies 40% of them as reconstructible, but potentially with low accuracy due to insufficient parallax, lack of non-repetitive texture, etc. The remaining 10% goes to the next stage. The VLM also extracts metadata such as whether the scene is dynamic or not.

Dynamic mask extraction. We use Grounding DINO [64] to detect bounding boxes of potentially movable object categories, such as people and cars. These regions are then excluded from matching, tracking, and verification.

Feature matching and tracking. We extract matches and tracks across frames with an ensemble of methods, using SIFT, SuperPoint + SuperGlue [21, 85], ALIKED + LightGlue [60, 141], and VGGSfM Tracker [114]. Matches within dynamic regions are discarded.

Reconstruction and filtering. We use VGGT to initialize camera parameters (when RANSAC-based essential matrix estimation yields too few inliers), and then run COLMAP [87] for iterative bundle adjustment and filtering based on the correspondences computed above. For successful reconstructions, we discard sequences that fail heuristic checks, *e.g.*, image registration ratio $< 99.5\%$, field of view outside $[30^\circ, 120^\circ]$, or distortion ratio > 0.1. These criteria aggressively remove cases with degenerate motion or extreme zoom. Then, we estimate per-frame dense depth maps using patch-based multi-view stereo [87].

Multi-view consistency. For each frame, we unproject the depth map to 3D, reproject the points into other views, and compare them with the depths there. Pixels that satisfy this cross-view consistency check are marked valid. We discard sequences with fewer than 5% valid depth pixels, which typically, though not always, indicates low-quality cameras.

Supervised geometric filtering. Finally, we have the cameras, depths, and valid masks for every sequence. We use handcrafted features, such as the camera up-vector consistency, parallax angle, or trajectory smoothness, to describe each sequence. We hand-annotated 500 static and 500 dynamic sequences, respectively, to train a classifier to remove low-quality reconstructions. The classifier is an ensemble of XGBoost [15], Random forests [10], and CatBoost [80].

4. Experiments

Here we introduce additional details of our framework, benchmark it against state-of-the-art rigid and dynamic re-

Table 1. **Camera Pose Estimation**, across static and dynamic benchmarks. The metric AUC is higher the better. Feed-forward models (DA3, PI3, VGGT) are robust across datasets, but still lag behind the dynamic optimization-based method MegaSaM at strict thresholds on Sintel (*e.g.*, AUC@3° of 16.2 *vs*. 22.5). In contrast, dynamic optimization methods (MegaSaM, MonST3R) degrade on wide-baseline scenes (*e.g.*, AUC@30° of 38.1 *vs*. 86.4 on ETH3D). Instead, our method substantially advances the state of the art across all the scenarios, *e.g.*, improving AUC@3° from 22.5 to 40.0 on Sintel, with a 77% relative improvement.

Method	Static Scenes						Dynamic Scenes					
	7 Scenes		NRGBD		ETH3D		DyCheck		Sintel		TUM-Dynamic	
	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°	AUC@3°	AUC@30°
MonST3R	9.0	68.3	13.9	79.7	1.7	14.3	11.5	45.4	4.3	45.8	7.7	48.5
MapAnything	5.8	61.4	35.2	88.9	13.2	51.0	6.1	60.3	2.9	31.6	4.3	40.2
MegaSaM	10.6	71.8	17.2	83.1	5.9	38.1	26.8	53.1	22.5	58.3	15.4	59.0
VGGT	10.9	74.4	81.7	97.7	18.8	62.1	21.0	78.7	15.0	50.0	16.6	61.2
PI3	13.3	77.0	83.8	98.2	35.3	79.6	23.3	81.0	14.8	53.5	16.1	59.2
DA3	18.7	78.2	86.4	98.4	46.1	87.0	32.1	83.9	16.2	52.7	20.8	62.7
Ours-1B	29.6	83.1	89.7	98.8	49.8	88.5	38.4	87.3	35.3	73.0	30.2	82.3
Ours-10B	**36.4**	**88.2**	**92.5**	**99.1**	**56.3**	**90.4**	**43.7**	**90.9**	**40.0**	**79.1**	**36.4**	**87.5**

Table 2. **Depth Estimation**, across static and dynamic benchmarks. The metric $\delta_{1.25}$ denotes the percentage of predicted depths within a factor of the ground truth (higher the better) and AbsRel is the mean absolute relative error (lower the better).

Method	Static Scenes						Dynamic Scenes					
	7 Scenes		NRGBD		ETH3D		DyCheck		Sintel		TUM-Dynamic	
	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel	$\delta_{1.25}$	AbsRel
MonST3R	92.4	0.075	98.4	0.030	95.8	0.056	93.3	0.068	71.9	0.263	85.0	0.148
MapAnything	92.9	0.070	98.7	0.022	96.3	0.035	97.0	0.049	72.5	0.251	93.1	0.052
MegaSaM	93.8	0.065	96.2	0.057	94.8	0.083	97.4	0.042	74.1	0.207	92.9	0.083
VGGT	91.9	0.073	99.1	0.019	97.4	0.036	95.2	0.055	79.2	0.189	92.2	0.064
PI3	92.8	0.068	99.2	0.011	99.6	0.016	97.4	0.041	82.5	0.144	95.5	0.046
DA3	93.0	0.063	99.5	0.010	99.6	0.015	97.7	0.039	86.1	0.118	94.3	0.049
Ours-1B	94.6	0.058	99.6	0.010	99.8	0.012	98.4	0.038	89.5	0.097	97.4	0.041
Ours-10B	**96.3**	**0.050**	**99.7**	**0.007**	**99.8**	**0.009**	**98.7**	**0.030**	**93.5**	**0.081**	**98.3**	**0.035**

construction methods, and ablate key design choices.

4.1. Implementation Details

VGGT-Ω includes four model variants with 200M, 500M, 1B, and 10B parameters, 12/12/24/16 alternating-attention blocks and hidden sizes 384/768/1024/4096, respectively. The vision transformer is initialized from DINOv3 [93] and is not frozen during training. Each block contains one global (or scene-token global) attention layer and one frame-wise attention layer. Optimization uses AdamW for 240K iterations, with 160K supervised, 50K self-supervised, and a final 30K supervised stage. The learning rate employs a linear warm-up over 5% of training and cosine decay for the remaining 95%, with a peak value of 2×10^{-4} for supervised and 1×10^{-4} for self-supervised. For each batch, the number of frames is drawn uniformly from $[1, 24]$. We augment images by randomizing the aspect ratio in $[0.33, 1.33]$ with the longer side fixed to 512 pixels and using color jittering, grayscale conversion, and random patch masking. Training was conducted on 128 96GB H100 GPUs using bfloat16 mixed precision, gradient checkpointing, and Fully Sharded Data Parallel (FSDP).

4.2. Benchmarking

We compare VGGT-Ω with recent approaches: (i) feed-forward reconstruction models (VGGT [115], PI3 [121], DA3 [59]) and (ii) optimization-based dynamic reconstruction methods (MonST3R [138], MegaSaM [57]). We evaluate on three static datasets (7 Scenes [92], NRGBD [5], and ETH3D [89]) and three dynamic datasets (DyCheck [30], Sintel [11], and TUM-Dynamic [97]). For each scene or sequence, we randomly sample 10 frames.

Following [115], we report the standard AUC for camera pose estimation (higher is better), computed as the area under the curve of the fraction of image pairs whose relative rotation and translation errors are below an angular threshold (*e.g.*, 3°, 30°). As shown in Tab. 1, feed-forward models generally exhibit strong performance on static benchmarks and at more relaxed thresholds, while optimization-based, dynamic-aware MegaSaM is more competitive on challenging dynamic sequences such as Sintel but degrades on wide-baseline or low-texture scenes. Our models generally outperform the baselines across both static and dynamic datasets and at both strict and relaxed thresholds.

We also evaluate the accuracy of the predicted depths with absolute relative error (AbsRel, lower is better) and $\delta_{1.25}$ (higher is better), which measures the percentage of pixels for which the ratio between the predicted depth and the ground-truth depth is within a specified threshold. As shown in Tab. 2, our models outperform the baselines across the static benchmarks, further lowering AbsRel on datasets where existing methods perform strongly, such as NRGBD and ETH3D, and delivering larger gains on dynamic scenes, where they reduce depth errors and increase $\delta_{1.25}$ (*e.g.*, on Sintel, improving $\delta_{1.25}$ from 86.1 to 93.5 and AbsRel from

0.118 to 0.081). The larger 10B variant consistently outperforms the 1B model, indicating that scaling up the reconstruction model directly benefits both camera and depth accuracy.

Additionally, we provide a qualitative visualization of our results in Fig. 4. It illustrates reconstructions on both static and dynamic scenes, including traffic, human motion, natural landscapes, and underwater environments.

4.3. Ablation Studies and Discussions

Unless otherwise specified, ablations use the 1B variant trained on 2M sequences with 64 GPUs for 150k supervised steps. To jointly assess camera and depth accuracy, we unproject depth maps into 3D using the cameras and compute the ℓ_2 distance between predicted and ground-truth points. We refer to this metric as point error. All the models are trained on approximately the same number of tokens.

Model and data size. We observe that scaling either the model or the data consistently improves performance, as shown in Fig. 1. Increasing the number of training sequences in $10\times$ steps yields a monotonic drop in point error, from 0.275 to 0.073. Overall, the shape of both curves suggests that power laws might characterize scaling in this class of models too.

Scene token global attention. A variant that uses only global attention layers achieves a point error of 0.071. Replacing 25% of the global attention layers with scene-token attention yields almost the same performance (0.073). However, increasing the fraction of scene-token layers beyond this point leads to a noticeable performance drop.

Multi-task learning. Removing the point and matching losses increases the point error from 0.073 to 0.078. For reference, adopting VGGT's original multi-head, multi-task setup achieves 0.070 but requires multiple dense heads, making scaling difficult.

Self-supervised training. Replacing 10% of training steps from supervised to self-supervised training slightly reduces point error from 0.073 to 0.070. We believe this stems from training on unlabeled data, which is more diverse. We also observe improved out-of-distribution generalization.

Annotation quality. To validate the quality of the pseudo ground truth produced by our annotation pipeline, we compare it against MegaSaM [57] on Sintel, which provides synthetic camera and depth ground truth. For a fair comparison, we evaluate only the sequences and pixels that satisfy both our filtering criteria and MegaSaM's validation, excluding $8/23$ sequences and all dynamic regions. Under this protocol, our pipeline achieves 96.4% AUC@30° for camera pose and 99.3% $\delta_{1.25}$ for depth, compared with 62.1% and 77.2% for MegaSaM, respectively, confirming the high quality of the resulting annotations. Our goal in pseudo-label generation is not to maximize yield, but to retain only sequences and pixels that are very likely to be correct, as we find that a smaller set of highly accurate pseudo ground truth is more beneficial than a larger but noisier collection in practice. Hence, the annotation pipeline is intentionally conservative: if a sequence is even mildly ambiguous, or if a pixel cannot be validated reliably, we prefer to discard it rather than risk injecting noisy supervision.

Table 3. **Performance on the LIBERO benchmark.** We freeze our pretrained model and feed the scene tokens as additional input to the OpenVLA-OFT and report success rate (SR) (higher is better). The clear performance gain validates the effectiveness of our scene tokens in aggregating spatial information.

Method	Spatial SR (%)	Object SR (%)	Goal SR (%)	Long SR (%)	Average SR (%)
Diffusion Policy	78.3	92.5	68.3	50.5	72.4
TraceVLA	84.6	85.2	75.1	54.1	74.8
Octo	78.9	85.7	84.6	51.1	75.1
OpenVLA	84.7	88.4	79.2	53.7	76.5
Dita	84.2	96.3	85.4	63.8	82.4
CoT-VLA	87.5	91.6	87.6	69.0	83.9
π_0-FAST	96.4	96.8	88.6	60.2	85.5
π_0	96.8	98.8	95.8	85.2	94.2
UniVLA	96.5	96.8	95.6	92.0	95.2
OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1
OpenVLA-OFT + Our Frozen Scene Tokens	**99.3**	**99.2**	**99.0**	**96.7**	**98.5**

5. Applications

VLA models like [1, 53] have explored leveraging reconstruction models to improve spatial understanding. Here we concatenate the original input tokens of OpenVLA-OFT [48] with the frozen scene tokens produced by our model, and train using the standard OpenVLA-OFT protocol. As shown in Table 3, the geometry-aware scene tokens consistently improve the performance across all the tasks in the LIBERO benchmark [62].

6. Conclusion

We presented VGGT-Ω, a feed-forward reconstruction model that achieves strong results across static and dynamic benchmarks. We improved the original VGGT in terms of architecture, data, and training by introducing scene-token attention, a single dense prediction head with multi-task losses, a large-scale annotation pipeline that handles dynamic content, and a self-supervised training protocol that leverages vast amounts of unlabeled videos. These ingredients allowed us to train our model at unprecedented scale and thus explore scaling in feed-forward 3D reconstruction. Our conclusion is that both model size and training data volume are crucial to achieve top performance, and that the latter is not yet saturated. Finally, we hope VGGT-Ω will be a useful foundation for the community to build on.

References

[1] Ali Abouzeid, Malak Mansour, Qinbo Sun, Zezhou Sun, and Dezhen Song. Geoaware-vla: Implicit geometry aware vision-language-action model. *arXiv preprint arXiv:2509.14117*, 2025. 8

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10), 2011. 2

[3] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulo, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *European Conference on Computer Vision*, pages 589–604. Springer, 2020. 6

[4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022. 6

[5] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 7

[6] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 6

[7] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: a synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proc. CVPR*, 2023. 6

[8] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *Proc. ECCV*, 2024. 2

[9] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3D shape from image streams. In *Proc. CVPR*, 2000. 3

[10] Leo Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001. 6

[11] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. ECCV*, 2012. 7

[12] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv*, 2001.10773, 2020. 6

[13] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *Proc. CVPR*, 2025. 3

[14] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proc. CVPR*, 2021. 2

[15] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 6

[16] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. *arXiv*, 2025. 3

[17] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025. 3

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. CVPR*, 2017. 6

[19] Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. SAIL-Recon: Large SfM by augmenting scene regression with localization. In *Proc. 3DV*, 2026. 3

[20] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: chunk it, loop it, align it—pushing vggt's limits on kilometer-scale long RGB sequences. *arXiv*, 2025. 3

[21] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: self-supervised interest point detection and description. In *Proc. CVPR Workshop*, 2018. 2, 6

[22] 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*, 2409.19152, 2024. 3

[23] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint description and detection of local features. In *Proc. CVPR*, 2019. 2

[24] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16774–16784, 2025. 3

[25] Olivier D. Faugeras and Stephen J. Maybank. Motion from point matches: Multiplicity of solutions. *IJCV*, 4(3), 1990. 2

[26] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: simultaneous 4D reconstruction and tracking in the world. In *Proc. ICCV*, 2025. 3

[27] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 6

[28] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building Rome on a cloudless day. In *Proc. ECCV*, 2010. 2

[29] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-Neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. In *Proc. NeurIPS*, 2022. 3

[30] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 7

[31] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *Proc. CVPR*, 2022. 6

[32] 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 *Proc. CVPR*, 2020. 3

[33] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. D^2USt3R: Enhancing 3D reconstruction with 4d pointmaps for dynamic scenes. *arXiv*, 2504.06264, 2025. 3

[34] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 1

[35] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004. 2

[36] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. *arXiv*, 2405.14868, 2024. 6

[37] Yuan-Ting Hu, Jiahong Wang, Raymond A Yeh, and Alexander G Schwing. Sail-vos 3d: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1418–1428, 2021. 6

[38] Guichen Huang, Ruoyu Wang, Xiangjun Gao, Che Sun, Yuwei Wu, Shenghua Gao, and Yunde Jia. LongSplat: Online generalizable 3D Gaussian splatting from long sequence images. In *Proc. ICCV*, 2025. 3

[39] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: video pose engine for 3D geometric perception. *arXiv*, 2508.10934, 2025. 3

[40] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[41] Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, et al. Megasynth: Scaling up 3d scene reconstruction with synthesized data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16441–16452, 2025. 6

[42] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: feed-forward 3D Gaussian Splatting from unconstrained views. *arXiv*, 2505.23716, 2025. 3

[43] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025. 3

[44] Dongki Jung, Jaehoon Choi, Yonghan Lee, Sungmin Eum, Heesung Kwon, and Dinesh Manocha. MoRe: Monocular geometry refinement via graph optimization for cross-view consistency. In *Proc. WACV*, 2026. 3

[45] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[46] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[47] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: universal feed-forward metric 3D reconstruction. *arXiv*, 2509.13414, 2025. 3

[48] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv*, 2502.19645, 2025. 8

[49] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *Proc. ICLR*, 2026. 3

[50] Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. Eden: Multimodal synthetic dataset of enclosed garden scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1579–1589, 2021. 6

[51] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MASt3R. In *Proc. ECCV*, 2024. 1

[52] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin,

Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation, 2024. 6

[53] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025. 8

[54] Samuel Li, Pujith Kachana, Prajwal Chidananda, Saurabh Nair, Yasutaka Furukawa, and Matthew Brown. Rig3R: Rig-aware conditioning for learned 3D reconstruction. *arXiv*, 2025. 3

[55] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*, 2018. 6

[56] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025. 2

[57] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: accurate, fast and robust structure and motion from casual dynamic videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 7, 8

[58] Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. WinT3R: Window-based streaming reconstruction with camera token pool. In *Proc. ICLR*, 2025. 3

[59] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 2, 7

[60] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: local feature matching at light speed. In *Proc. ICCV*, 2023. 2, 6

[61] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: a large-scale scene dataset for deep learning-based 3d vision. In *Proc. CVPR*, 2024. 6

[62] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In *Proc. NeurIPS*, 2023. 8

[63] Changkun Liu, Bin Tan, Zeran Ke, Shangzhan Zhang, Jiachen Liu, Ming Qian, Nan Xue, Yujun Shen, and Tristan Braud. PLANA3R: Zero-shot metric planar 3D reconstruction via feed-forward planar splatting. *arXiv*, 2025. 3

[64] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6

[65] Shaohui Liu, Yidan Gao, Tianyi Zhang, Rémi Pautrat, Johannes L Schönberger, Viktor Larsson, and Marc Pollefeys. Robust incremental structure-from-motion with hybrid features. In *Proc. ECCV*, 2024. 2

[66] Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. uCO3D uncommon objects in 3D. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6

[67] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3R: Aligned monocular depth estimation for dynamic videos. In *Proc. CVPR*, 2025. 3

[68] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar RAFT. *arXiv.cs*, abs/2205.04502, 2022. 3

[69] Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the $sl(4)$ manifold. In *Proc. NeurIPS*, 2025. 3

[70] Soroush Mahdi, Fardin Ayar, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. Evict3R: Training-free token eviction for memory-bounded streaming visual geometry transformers. *arXiv*, 2025. 3

[71] Mapillary. Metropolis dataset. https://www.mapillary.com/dataset/metropolis. Accessed 2025-11-09. 6

[72] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *Proc. CVPR*, 2020. 3

[73] John Oliensis. A critique of structure-from-motion algorithms. *CVIU*, 80(2), 2000. 2

[74] 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, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 5

[75] Onur Özyesil, V. Voroninski, R. Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26, 2017. 2

[76] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Con-*

ference on Computer Vision (ICCV), pages 20133–20143, 2023. 6

[77] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *Proc. ICCV*, 2023. 6

[78] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 20524–20531. IEEE, 2025. 6

[79] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proc. CVPR*, 2022. 3

[80] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018. 6

[81] Dexin Qi, Tao Tao, Zhihong Zhang, and Xuesong Mei. Fupad: Scalable pose estimation by fusing patch-wise vggt with dense bundle adjustment. In *Proc. ICIRA*, 2025. 3

[82] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. ICCV*, 2021. 4

[83] 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 *Proc. ICCV*, 2021. 6

[84] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proc. ICCV*, 2021. 6

[85] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: learning feature matching with graph neural networks. In *Proc. CVPR*, 2020. 2, 6

[86] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proc. ICCV*, 2019. 6

[87] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 1, 4, 6

[88] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. ECCV*, 2016. 2

[89] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 7

[90] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 2, 4

[91] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. ClusterGNN: cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proc. CVPR*, 2022. 2

[92] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 7

[93] 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*, 2025. 1, 2, 4, 5, 7

[94] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. on Graphics (TOG)*, 2006. 1, 2

[95] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv*, 1906.05797, 2019. 6

[96] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models. *arXiv preprint arXiv:2406.10224*, 2024. 6

[97] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 7

[98] Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic point maps: A versatile representation for dynamic 3d reconstruction. *arXiv preprint arXiv:2503.16318*, 2025. 3

[99] Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic Point Maps: A versatile representation for dynamic 3D reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025. 5

[100] Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: Video reconstruction with dynamic point maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 5

[101] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 6

[102] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment network. In *Proc. ICLR*, 2019. 2

[103] 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 *Proc. CVPR*, 2025. 3

[104] Zachary Teed and Jia Deng. DeepV2D: video to depth with differentiable structure from motion. In *Proc. ICLR*, 2020. 2

[105] Zachary Teed and Jia Deng. DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Proc. NeurIPS*, 2021. 2

[106] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 30(5), 2008. 3

[107] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8942–8952, 2021. 6

[108] MJ Tyszkiewicz, P Fua, and E Trulls. DISK: learning local features with policy gradient. In *Proc. NeurIPS*, 2020. 2

[109] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: depth and motion network for learning monocular stereo. In *Proc. CVPR*, 2017. 2

[110] Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster vggt with block-sparse global attention. *arXiv preprint arXiv:2509.07120*, 2025. 2, 4

[111] Hengyi Wang and Lourdes Agapito. Spann3R: 3D reconstruction with spatial memory. In *Proc. 3DV*, 2024. 3

[112] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *Proc. CVPR*, 2021. 2

[113] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: solving pose estimation via diffusion-aided bundle adjustment. In *Proc. ICCV*, 2023.

[114] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: visual geometry grounded deep structure from motion. In *Proc. CVPR*, 2024. 3, 4, 6

[115] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 3, 4, 7

[116] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiaoxiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: a large-scale video dataset with spatial annotations. *arXiv*, 2509.09676, 2025. 2

[117] Lijie Wang, Lianjie Guo, Ziyi Xu, Qianhao Wang, Fei Gao, and Xieyuanli Chen. LiDAR-VGGT: Cross-modal coarse-to-fine fusion for globally consistent and metric-scale dense mapping. *arXiv*, 2025. 3

[118] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. *arXiv*, 2501.12387, 2025. 3

[119] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proc. CVPR*, 2024. 1, 3, 5

[120] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: a dataset to push the limits of visual SLAM. In *Proc.IROS*, 2020. 6

[121] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv*, 2507.13347, 2025. 1, 3, 7

[122] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. DeepSFM: structure from motion via deep bundle adjustment. In *Proc. ECCV*, 2020. 3

[123] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proc. ICCV*, 2021. 3

[124] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Proc. NeurIPS*, 2025. 3

[125] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024. 6

[126] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: tracking any 2d pixels in 3d space. *arXiv*, 2404.04319, 2024. 3

[127] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. *arXiv preprint arXiv:2507.12462*, 2025. 3

[128] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D point tracking made easy. In *Proc. ICCV*, 2025. 3

[129] 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. *Proc. CVPR*, 2025. 3

[130] David Yifan Yao, Albert J Zhai, and Shenlong Wang. Uni4d: Unifying visual foundation models for 4d modeling from a single video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1116–1126, 2025. 3

[131] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proc. ECCV*, 2018. 3

[132] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Proc. NeurIPS*, 2020. 3

[133] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: a high-fidelity dataset of 3d indoor scenes. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 6

[134] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: learned invariant feature transform. In *Proc. ECCV*, 2016. 2

[135] Yuheng Yuan, Qiuhong Shen, Shizun Wang, Xingyi Yang, and Xinchao Wang. Test3R: Learning to reconstruct 3D at test time. In *Proc. NeurIPS*, 2025. 3

[136] Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. CVPR*, 2018. 6

[137] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 1

[138] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: a simple approach for estimating geometry in the presence of motion. *arXiv*, 2410.03825, 2024. 3, 5, 7

[139] 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 *Proc. CVPR*, 2025. 3

[140] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. GeoMVSNet: Learning multi-view stereo with geometry perception. In *Proc. CVPR*, 2023. 3

[141] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. ALIKED: a lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Trans. on Instrumentation and Measurement*, 72, 2023. 6

[142] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proc. CVPR*, 2023. 6

[143] Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. PAGE-4D: disentangled pose and geometry estimation for 4D perception. *arXiv*, 2510.17568, 2025. 3

[144] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. CVPR*, 2017. 3

[145] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *Proc. ICLR*, 2026. 3