MVGE: SCALE-INVARIANT AND TEMPORAL-CONSISTENT MONOCULAR VIDEO GEOMETRY ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present MVGE, a novel approach for estimating 3D geometry from extended monocular video sequences, where existing methods struggle to maintain both geometric accuracy and temporal consistency across hundreds of frames. Our approach generates affine-invariant 3D point maps with shared parameters across entire sequences, enabling consistent scale-invariant representations. We introduce three key innovations: viewpoint-invariant geometry aligning multi-perspective points in a unified reference frame; appearance-invariant learning enforcing consistency across exponential timescales; and frequency-modulated positioning enabling extrapolation to sequences vastly exceeding training length. Experiments across diverse datasets demonstrate significant improvements, reducing relative point map error by 24.2% and temporal alignment error by 34.9% on ScanNet compared to state-of-the-art methods. Our approach handles challenging scenarios with complex camera trajectories and lighting variations while efficiently processing extended sequences in a single pass. Code will be publicly released, and we encourage readers to explore the interactive demonstrations in our supplementary materials.

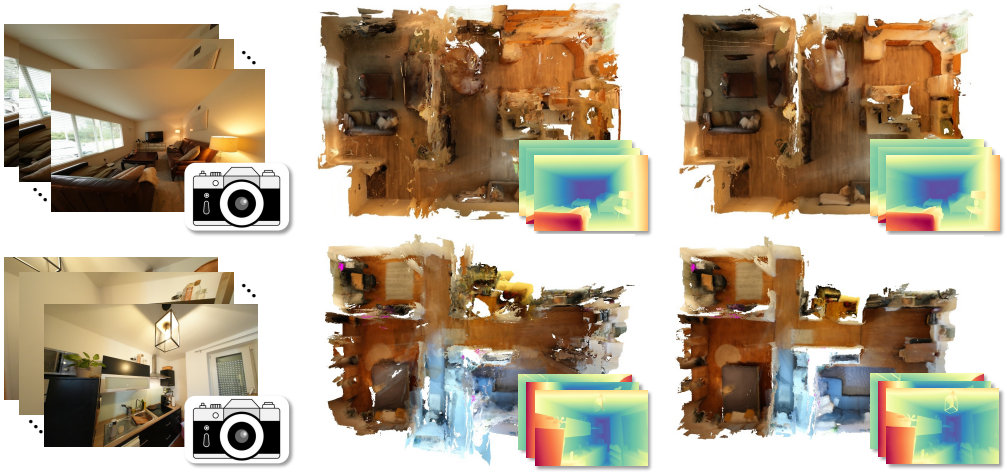

Figure 1: Given a sequence of video frames, MVGE is capable of predicting scale-invariant and temporal-consistent point maps in a single forward pass. We visualize the 3D mesh reconstructed by TSDF integration of 100 point maps predicted by MVGE in a single shot, in comparison with MoGe (Wang et al., 2025b) using ScanNet++ (Yeshwanth et al., 2023) dataset. MVGE maintains geometric accuracy and long-range consistency across hundreds of frames with minimum drift, enabling high-quality 3D reconstruction.

1 INTRODUCTION

Estimating 3D geometry from monocular videos is a fundamental challenge in computer vision with diverse applications in novel view synthesis, autonomous navigation, virtual reality, and 3D/4D reconstruction. Despite significant advances in single-image depth estimation, video-based approaches still struggle with two critical challenges: achieving **high geometric accuracy at multiple scales** within each frame and across the global coordinate system, while maintaining **temporal consistency throughout sequences of hundreds of frames without scale drift**.

Existing methods typically excel in one area at the expense of the other. Single-image approaches like MoGe (Wang et al., 2025b) capture detailed geometry but produce inconsistent results when applied frame-by-frame to videos. Conversely, video-specific methods (Yang et al., 2025; Hu et al., 2025; Chen et al., 2025; Wang et al., 2024a; Zhang et al., 2025) inherently lack geometric precision while providing only short-term consistency, still exhibiting significant scale drift in longer sequences. Traditional approaches rely on optical flow constraints (Wang et al., 2023; 2024b) that only link adjacent frames, failing to prevent accumulation of errors. Video diffusion models offer consistency through learned priors but at significant computational cost. Recent transformer-based approaches like VGGT (Wang et al., 2025a) can process longer sequences but lack effective temporal position encoding, limiting their effectiveness with complex camera motions.

Processing lengthy sequences with **geometric and temporal accuracy** requires **simultaneous consideration of hundreds of frames** with **precise temporal position encoding** to handle complex scene transformations. However, memory constraints make training on such long sequences impractical. This creates a fundamental tension: models need to **extrapolate effectively** to sequences far longer than their training examples. With robust extrapolation capabilities, overlapping inference techniques can achieve minimal drift by maintaining substantial frame overlap between consecutive windows, effectively scaling to unlimited sequence lengths.

We present a novel approach generating **affine-invariant** 3D point maps from RGB videos with both geometric precision and long-range temporal consistency. Our method produces point maps where all frames share the same scale and shift parameters, with a unified optimization approach to recover scale-invariant representations for downstream applications. Our key innovations include: **Viewpoint-invariant geometry** transforming points from multiple perspectives into a shared reference frame through camera pose integration; **Appearance-invariant learning** that supervises geometric consistency across exponential time scales while isolating persistent structural features from transient visual conditions; and **Adaptive frequency-modulated positioning** implementing an NTK-guided rotary scheme with strategic training-time extrapolation simulation to process sequences orders of magnitude longer than training examples. Our approach significantly outperforms previous methods, reducing relative point map error by 24.2% on ScanNet and temporal alignment error by 34.9% compared to existing approaches, while maintaining superior performance across diverse datasets from synthetic animations to real-world driving scenarios.

2 RELATED WORK

Monocular depth estimation. Recent advances in monocular depth estimation have significantly improved both geometric accuracy and generalization. Early supervised approaches (Eigen et al., 2014; Fu et al., 2018; Bhat et al., 2021; 2023) were limited by domain-specific datasets. More recent methods overcame this limitation through affine-invariant representations (Ranftl et al., 2022; Birkl et al., 2023; Ranftl et al., 2021) or scale alignment techniques (Yin et al., 2023; Hu et al., 2024). Large-scale data-driven approaches (Yang et al., 2024a;b) and diffusion-based models (Ke et al., 2024; Gui et al., 2024; Fu et al., 2024) have further enhanced generalization to diverse scenarios. While some methods (Yin et al., 2021b; Piccinelli et al., 2024; 2025; Bochkovskii et al., 2025) predict both depth and camera intrinsics, they often lack precision in local geometry. MoGe (Wang et al., 2025b) achieves superior geometric accuracy through multi-scale supervision but operates only on single images, lacking cross-frame consistency.

Video-based depth estimation. Extending depth estimation to video sequences introduces significant temporal consistency challenges. Video diffusion models (Hu et al., 2025; Yang et al., 2025)

provide inherent coherence but at high computational cost. For sequences longer than training examples, several strategies have emerged: sliding windows (Hu et al., 2025; Chen et al., 2025), keyframe conditioning (Yang et al., 2025), and global attention (Wang et al., 2025a). However, these methods still exhibit scale drift over extended sequences or struggle with complex camera trajectories. Current approaches typically excel at either geometric accuracy or temporal consistency, rarely achieving both across hundreds of frames.

Positional encoding for extrapolation. Transformers struggle with sequences longer than their training examples. While standard sinusoidal encodings (Vaswani et al., 2017) and learned embeddings have limited extrapolation capabilities, Rotary Position Encoding (RoPE) (Su et al., 2021) better generalizes by encoding relative positions through complex plane rotations. Strategic frequency adjustments (Chen et al., 2023; Peng et al., 2024) and NTK-aware adaptations (Peng & Quesnelle, 2023; Sun et al., 2022) preserve both local details and global structure during extrapolation. Our work adapts these techniques, primarily developed for language models, to video-based 3D reconstruction, enabling effective processing of sequences substantially longer than training examples.

3 METHOD

We present a novel approach for generating geometrically accurate and temporally consistent 3D point maps from RGB videos. Our method addresses two critical challenges: producing geometrically precise representations for each frame, and maintaining long-range temporal consistency across hundreds of frames - essential requirements for downstream 3D reconstruction tasks.

3.1 GEOMETRY-AWARE VIDEO POINT MAP ESTIMATION

Task definition. Given an RGB video sequence $\mathbf{I} = \{I_1, I_2, ..., I_T\}$ with T frames, our goal is to estimate scale-invariant and temporally consistent 3D point maps from unposed monocular videos. Specifically, we predict a sequence of 3D point maps $\mathbf{P} = \{P_1, P_2, ..., P_T\}$, where $P_t \in \mathbb{R}^{H\times W\times 3}$ represents the 3D coordinates of each pixel in frame t within that frame's camera coordinate system. **Training setup:** During training, our method takes multi-frame RGB images as network input and optionally uses ground truth camera poses solely for computing the cross-frame geometric loss. The poses enable multi-scale geometric supervision by transforming predicted point clouds to a common reference frame, but are not required for all training data. **Inference setup:** At inference, our method requires only multi-frame RGB images as input and outputs scale-consistent point maps in each frame's camera coordinate system. These point maps can subsequently serve as input to methods like MegaSAM (Li et al., 2025) to estimate camera parameters and enable high-quality 4D reconstruction.

Positioning relative to existing approaches. Our method addresses fundamental limitations of existing approaches across three categories: **Single-frame pointmap methods** (e.g., DepthPro (Bochkovskii et al., 2025), MoGe (Wang et al., 2025b)) process frames independently, leading to scale inconsistencies that degrade downstream reconstruction quality. Our approach achieves superior long-range temporal consistency and global geometric accuracy, enabling more precise camera pose estimation and 4D reconstruction. **Video depth methods** (e.g., DepthCrafter (Hu et al., 2025), Video Depth Anything (Chen et al., 2025)) typically output affine-invariant depth maps, where the missing shift parameter and camera intrinsics complicate direct 4D reconstruction. Compared to these video depth prediction methods, our approach maintains scale consistency across significantly longer temporal sequences. **Single coordinate system methods** (e.g., Dust3r (Wang et al., 2024a), MonST3R (Zhang et al., 2025)) directly estimate global point maps and camera poses jointly. While our approach requires external pose estimation, it produces more accurate 4D reconstructions and handles significantly longer video sequences under identical memory constraints.

Scale-Invariant representation. Our model predicts affine-invariant point maps following MoGe (Wang et al., 2025b), where each point map is agnostic to global scale $s \in \mathbb{R}$ and offset $\mathbf{t} \in \mathbb{R}^3$. The key distinction is that our entire video sequence shares these parameters: $P_i \cong sP_i + \mathbf{t}, \forall i \in [1, T]$. During inference, we recover a single shared focal length f and Z-axis shift t_z for all frames by minimizing the projection error:

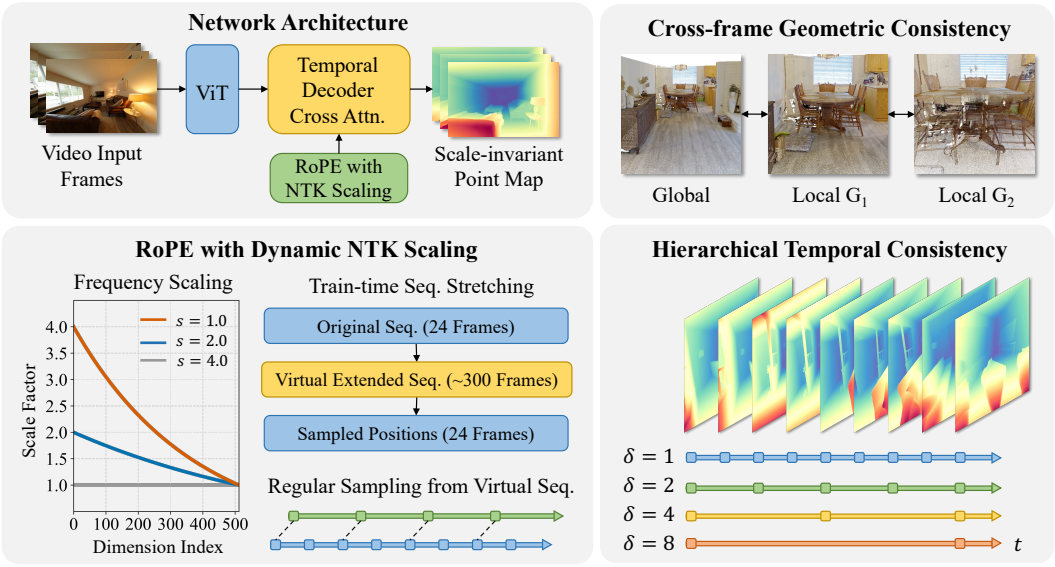

Figure 2: **Overview of MVGE.** *Top-Left:* MVGE consists of a ViT backbone that processes video input frames, followed by a temporal decoder with cross-attention and dynamic NTK scaling RoPE, producing scale-invariant point maps (Sec. 3.1). *Top-Right:* Cross-frame geometric consistency enforced across global and local geometric levels (G_1, G_2) to maintain structural coherence across frames (Sec. 3.1). *Bottom-Left:* RoPE with dynamic NTK scaling applied to extend sequence context, using frequency scaling that adaptively weights dimensions based on scale factor, and train-time sequence stretching that creates a virtual extended sequence to sample positions (Sec. 3.2). *Bottom-Right:* Hierarchical temporal consistency constraints applied multiple temporal strides ($\delta = 1, 2, 4, 8$) to enforce smooth, consistent point map predictions across time (Sec. 3.2).

$$\min_{f,t_z} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(\frac{f x_{t,i}}{z_{t,i} + t_z} - u_{t,i} \right)^2 + \left(\frac{f y_{t,i}}{z_{t,i} + t_z} - v_{t,i} \right)^2, \tag{1}$$

where $(x_{t,i}, y_{t,i}, z_{t,i})$ are the predicted 3D coordinates and $(u_{t,i}, v_{t,i})$ are the corresponding 2D pixel coordinates. This ensures a metrically consistent representation across the entire video, essential for 3D reconstruction tasks. By recovering the shift parameter, we transform our predictions into scale-invariant point maps, making them directly applicable for downstream tasks such as 3D reconstruction and novel view synthesis.

Geometric accuracy through multi-scale training. To achieve fine-grained geometric accuracy, we adopt MoGe's multi-scale approach (Wang et al., 2025b). MoGe achieves superior geometric precision through three key mechanisms: (1) affine-invariant alignment that handles ambiguities in depth perception, (2) multi-scale local geometry supervision that enforces accuracy at different spatial scales, and (3) normal consistency loss that ensures surface coherence. These mechanisms collectively enable the capture of both global structure and fine geometric details.

Building on this foundation, our approach extends the geometric accuracy requirements to video sequences. Additionally, other works (Bochkovskii et al., 2025; Yang et al., 2024b; Yin et al., 2021a; 2019) have demonstrated that spatial gradient regularization can further improve detailed depth estimation by constraining the local surface structure. We incorporate these principles into our video-based framework to maintain high-fidelity geometric representations across frames.

Cross-frame geometric constraints. To enforce geometric consistency across frames, we transform all points to a common reference frame using camera poses. We randomly select one frame from the sequence as the reference frame for each training iteration, providing diverse viewpoints during training. This process involves first converting points from individual camera coordinates to world coordinates using the camera-to-world transforms, and then transforming these world points to the

randomly selected reference frame. This transformation allows us to directly compare geometric structures captured from different viewpoints within a unified coordinate system.

We then apply a multi-scale geometric loss framework to the points in this common reference frame:

$$\mathcal{L}_{cross} = \sum_{l \in \{1, G_1, G_2\}} \frac{1}{|C_l|} \sum_{c \in C_l} \frac{1}{|M_c|} \sum_{i \in M_c} w_i \cdot \| s_c \cdot \mathbf{p}_{pred}^{ref}[i] + \mathbf{t}_c - \mathbf{p}_{gt}^{ref}[i] \|_1, \tag{2}$$

where l is the grid size (with $l = 1$ representing global alignment), C_l is the set of cells at grid size l, M_c is the set of valid points in cell c, w_i is a depth-aware weight, and $(s_c, \mathbf{t}_c)$ are alignment parameters computed independently for each cell.

For global alignment ($l = 1$), the entire point cloud is treated as a single cell. For local alignment, we divide the 3D space into a grid of $G_l \times G_l \times G_l$ cells. Our implementation uses grid sizes of 4 and 16, allowing the model to capture both coarse structure and fine details across the entire temporal sequence. By enforcing geometric consistency at multiple scales, our approach ensures that the predicted point maps maintain both local detail and global structure across the video.

3.2 Long-Range Temporal Consistency

Temporal consistency challenges. Downstream reconstruction tasks require point maps that exhibit: (1) consistent geometric accuracy both within individual frames and across the entire sequence at local and global scales, and (2) temporal stability over extended sequences rather than just between adjacent frames. When these requirements aren't met, particularly under challenging conditions with dramatic lighting changes or significant camera movements, scale drift can accumulate, severely degrading reconstruction quality and producing distorted or fragmented results.

Recent video diffusion model-based approaches (Hu et al., 2025; Yang et al., 2025; Shao et al., 2025) leverage inherent temporal consistency mechanisms, but suffer from significant computational inefficiency. Other methods utilize optical flow-based losses (Wang et al., 2023; 2024b; Kuang et al., 2025; Chen et al., 2025) to maintain consistency between adjacent frames. However, these approaches only constrain relationships between consecutive frames, causing error accumulation over longer sequences. Furthermore, they struggle with large camera motions, which can substantially degrade depth prediction accuracy by introducing conflicting geometric constraints when camera viewpoint changes significantly.

Structure-Preserving temporal supervision. To address fundamental limitations in temporal consistency, we introduce a hierarchical derivative supervision framework that operates across multiple time scales:

$$\mathcal{L}_{temp} = \sum_{s=0}^{S-1} \frac{1}{|M_s|} \sum_{t=1}^{T-\delta_s} \sum_{i \in \mathcal{M}_{t,t+\delta_s}} w_{t,i} \cdot \left| \frac{\partial D_{pred}}{\partial t}(t,i) - \frac{\partial D_{gt}}{\partial t}(t,i) \right|, \tag{3}$$

where s indexes temporal scale, S is the total number of scales, $\delta_s = 2^s$ represents exponentially increasing time intervals, T is the sequence length, $\mathcal{M}_{t,t+\delta_s}$ denotes valid corresponding pixels between frames t and $t + \delta_s$, $|M_s|$ is the total number of valid pixels at scale s, $w_{t,i}$ is a depth-aware weight for pixel i in frame t, and $\frac{\partial D}{\partial t}$ represents the temporal derivative of depth values.

To disentangle geometric structure from appearance variations, we apply frame-specific augmentations with independently sampled color transformations and blur patterns across the sequence. This forces the model to focus on invariant geometric features while ignoring transient visual cues, enabling robust geometric consistency even under dramatic lighting changes and complex camera movements that typically challenge conventional methods.

Scaling beyond memory constraints. Processing hundreds of frames simultaneously during training is infeasible due to memory constraints. Existing methods address this limitation in various ways: DepthCrafter (Hu et al., 2025) and Video Depth Anything (Chen et al., 2025) use overlapping frame windows during inference, but suffer from scale drift due to limited training sequence length. Depth Anything Video (Yang et al., 2025) processes key frames first and then uses them as conditions for

other frames, but this approach has limited scalability and reduced efficiency. VGGT (Wang et al., 2025a) employs global attention without temporal information injection, struggling with complex camera trajectories where temporal relationships are critical.

Frequency-modulated extrapolation. To ensure robust handling of complex spatial relationships while enabling effective extrapolation to sequences much longer than those seen during training, we employ a specialized Rotary Position Encoding (RoPE) (Su et al., 2021; Chen et al., 2023; Peng & Quesnelle, 2023; Sun et al., 2022) with Neural Tangent Kernel (NTK) adaptation.

Our implementation computes frequency components with dynamic NTK scaling:

$$\theta_{i,j} = \frac{j \cdot s^{(1-\frac{i}{d})}}{10000^{\frac{2i}{d}}}, \tag{4}$$

where $\theta_{i,j}$ is the rotation angle, j is the position index, i indexes the frequency dimension, d is the embedding dimension, and $s = \frac{L_{seq}}{L_{train}}$ is a scaling factor applied when inference sequence length exceeds training length. This adaptive scaling preserves the model's capacity to capture both fine-grained temporal patterns and global structure by applying graduated adjustments across the frequency spectrum—attenuating changes to high-frequency components that encode local details while amplifying adjustments to low-frequency components that capture long-range dependencies.

During training, we randomly apply sequence stretching with 50% probability, where we generate position encodings for a virtual extended sequence and sample them at appropriate intervals to match the original sequence length. Mathematically, this involves computing $\theta'_{i,j}$ for a virtual sequence of length $L_{virtual} = L_{seq} \cdot r$ (where r is randomly sampled) and then sampling positions $j' = j \cdot r$ to obtain the final encodings. This technique simulates extrapolation during training, teaching the model to handle sequences significantly longer than those in the training data.

To further enhance temporal generalization, we employ variable temporal context windows during training. While maintaining a fixed 24-frame input size, we dynamically adjust the temporal stride between frames, allowing these 24 frames to represent contexts spanning from densely sampled short sequences to sparsely sampled long sequences of several hundred frames. This adaptive sampling strategy complements our position encoding approach, enabling the model to simultaneously learn representations for both fine-grained frame-to-frame transitions and long-range temporal relationships.

4 Experiment

4.1 Implementation Details

Model architecture. Our model builds upon MoGe (Wang et al., 2025b) by integrating temporal modeling capabilities through strategically placed temporal attention modules in the decoder. Specifically, we insert four transformer-based temporal modules after each feature level of the decoder with 8 attention heads and single-block architecture, enabling effective information exchange across frames while preserving spatial details. We employ DINOv2-L (Oquab et al., 2023) as our visual encoder and initialize all parameters from MoGe's pretrained weights.

Training datasets. For training, we use a diverse collection of synthetic datasets including TartanAir (Wang et al., 2020), PointOdyssey (Zheng et al., 2023), SPRING (Mehl et al., 2023), VKitti2 (Cabon et al., 2020), Lightwheel (LightwheelAI & contributors, 2024), Hypersim (Roberts et al., 2021), GTAIM (Cao et al., 2020), MVSSynth (Huang et al., 2018), UnrealStereo4K (Tosi et al., 2021), GTASFM (Wang & Shen, 2019), IRS (Wang et al., 2021), and MidAir (Fonder & Droogenbroeck, 2019).

Optimization strategy. We optimize using AdamW following MoGe's base configuration with learning rates of 10^{-4} for decoder parameters and 10^{-5} for encoder parameters. These rates are dynamically scaled according to the square root of batch size ratio, using a reference batch size of 32 frames as baseline. Our learning schedule employs warmup, linear decay, and step decay phases with all milestone parameters proportionally adjusted based on total iteration count. Throughout training, we preserve input aspect ratios while resizing images to maintain spatial relationships in the scene.

Loss functions. Our loss function integrates MoGe's original components with additional spatial and temporal consistency objectives. We maintain the affine-invariant global loss (weight 1.0), multi-scale

Method	Sintel		ScanNet			Bonn		KITTI		Avg.
	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$	$\text{TAE}^p\downarrow$	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$	Rank↓
DepthPro (Bochkovskii et al., 2025)	0.400	0.441	0.132	0.942	0.095	0.130	0.975	0.191	0.810	3.67
VGGT (Wang et al., 2025a)	0.382	**0.694**	0.032	0.992	0.079	**0.043**	**0.987**	0.196	0.764	2.67
MoGe (Wang et al., 2025b)	0.281	0.627	0.132	0.896	0.126	0.086	0.967	0.101	0.971	2.25
Ours	**0.257**	0.617	**0.100**	**0.961**	**0.082**	0.068	0.979	**0.091**	**0.976**	**1.25**
	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{TAE}^d\downarrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	Rank↓
DepthPro (Bochkovskii et al., 2025)	0.363	0.476	0.089	0.929	0.065	0.056	0.973	0.092	0.912	3.33
VGGT (Wang et al., 2025a)	0.359	**0.680**	0.029	0.989	0.048	**0.040**	**0.981**	0.187	0.728	2.67
MoGe (Wang et al., 2025b)	0.255	0.603	0.130	0.852	0.077	0.081	0.959	0.087	0.958	2.50
Ours	**0.216**	0.648	**0.081**	**0.941**	**0.049**	0.055	0.971	**0.081**	**0.965**	**1.25**

Table 1: **Evaluation on point map estimation and depth estimation.** Results are aligned with the ground truth by optimizing a shared scale factor across the entire video. Lower values are better for Rel and TAE (↓), while higher values are better for δ (↑). The best results in each column are highlighted in **bold**. Gray values indicate methods trained on ScanNet.

Pos. Encoding	Sintel			ScanNet			Bonn	
	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$	$\text{TAE}^p\downarrow$	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$	$\text{TAE}^p\downarrow$	$\text{Rel}^p\downarrow$	$\delta^p\uparrow$
None	0.304	**0.503**	0.426	0.163	0.878	0.089	0.118	0.958
APE	0.324	0.475	0.451	0.153	0.895	0.089	0.115	0.956
RoPE	0.307	0.491	0.410	0.140	0.915	0.092	0.103	**0.964**
RoPE+	**0.304**	**0.503**	**0.394**	**0.138**	**0.923**	**0.086**	**0.095**	0.963
Pos. Encoding	Sintel			ScanNet			Bonn	
	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{TAE}^d\downarrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{TAE}^d\downarrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$
None	0.261	0.547	0.246	0.107	0.896	0.053	0.073	0.954
APE	0.279	0.526	0.252	0.108	0.889	0.052	0.076	0.947
RoPE	0.261	0.547	0.235	0.097	0.911	0.055	**0.062**	**0.959**
RoPE+	**0.253**	**0.563**	**0.222**	**0.095**	**0.919**	**0.047**	0.064	**0.959**

Table 2: **Ablation study on extrapolation strategies.** Position encoding methods on 270-frame sequences exceeding our 24-frame training sequences. RoPE+ combines NTK-adapted rotary encoding with sequence stretching training.

Method	Sintel		Bonn		FPS
	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	
DepthCrafter	0.30	0.70	0.13	0.85	0.94
Video Depth Any.	0.30	0.64	0.07	0.96	4.47
DepthAnyVideo	0.41	0.66	**0.06**	**0.97**	6.48
MVGE (Ours)	**0.20**	**0.73**	**0.06**	**0.97**	**39.1**
Method	ScanNet		KITTI		Time (s)
	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	$\text{Rel}^d\downarrow$	$\delta^d\uparrow$	
DepthCrafter	0.17	0.73	0.15	0.77	320.1
Video Depth Any.	**0.09**	0.92	0.08	0.95	67.2
DepthAnyVideo	**0.09**	**0.93**	0.11	0.89	46.3
MVGE (Ours)	**0.09**	**0.93**	**0.07**	**0.97**	**7.7**

Table 3: **Video depth methods comparison.** Evaluation on 300 frames at 378×672 resolution with affine-invariant alignment.

local losses at levels 4, 16, and 64 (weights 1.0 each), normal loss (1.0), and mask loss (1.0). We adopt established spatial gradient loss (4.0) to preserve depth details, and introduce our proposed $\mathcal{L}_{temp}$ (2.0) and $\mathcal{L}_{cross}$ (1.0). For frame-specific augmentation, we apply color jitter and Gaussian blur with 0.5 probability to enhance robustness to appearance variations.

Computational resources. We trained our final model on 16 NVIDIA H20 GPUs for approximately 4.3 days. Each ablation study experiment required approximately 0.6 days of training on the same hardware configuration.

4.2 Evaluation

Evaluation datasets. We evaluate on five diverse datasets spanning various scenarios: **Sintel** (Butler et al., 2012) consists of 23 synthetic videos with 50 frames each, providing precise depth labels in complex scenes with challenging lighting and motion. **ScanNet v2** (Dai et al., 2017) includes 100 indoor test videos with rich geometric structures, from which we sample every third frame to create 90-frame sequences for standard evaluation. **Bonn** (Palazzolo et al., 2019) contains 26 dynamic videos with prominent foreground motions, where we use frames 30-140 to assess robustness to object movement. **KITTI** (Geiger et al., 2013) provides 13 outdoor driving sequences, from which we use the first 110 frames per sequence from the full validation set to evaluate performance in structured environments. **DDAD** (Guizilini et al., 2020) is an autonomous driving dataset featuring diverse outdoor scenes captured across varying weather conditions and environments, with sequences ranging from 50 to 100 frames. For ablation studies focusing on long-range temporal consistency, we extend our evaluation to 270-frame sequences using consistent sampling strategies across datasets where ground truth is available.

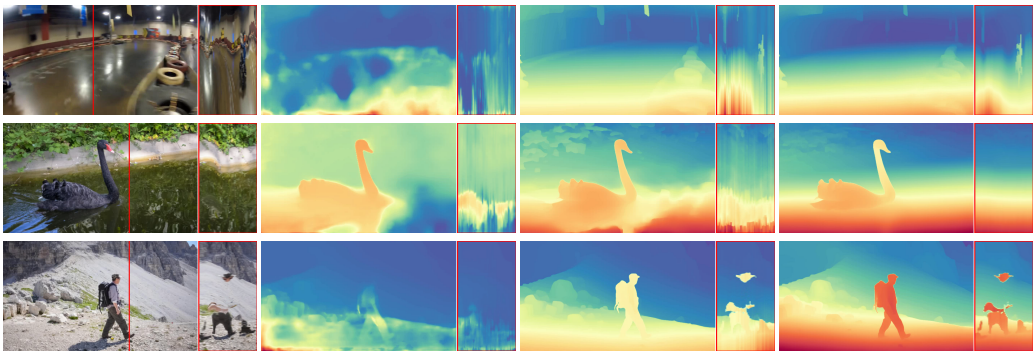

Figure 3: **Qualitative visualizations of depth predictions across diverse scenarios.** Each row shows an input frame with its corresponding spacetime slice (right portion), comparing depth predictions from VGGT, MoGe, and our method.

Inference Method	ScanNet (270 frames)				KITTI (270 frames)				DDAD			
	$\text{Rel}^p \downarrow$	$\delta^p \uparrow$	$\text{Rel}^d \downarrow$	$\delta^d \uparrow$	$\text{Rel}^p \downarrow$	$\delta^p \uparrow$	$\text{Rel}^d \downarrow$	$\delta^d \uparrow$	$\text{Rel}^p \downarrow$	$\delta^p \uparrow$	$\text{Rel}^d \downarrow$	$\delta^d \uparrow$
Sliding Window	0.114	0.935	0.098	0.908	0.102	0.963	0.097	0.930	0.192	0.863	0.115	0.894
Single-Pass	**0.113**	**0.937**	**0.094**	**0.913**	**0.092**	**0.974**	**0.084**	**0.963**	**0.187**	**0.879**	**0.108**	**0.916**

Table 4: **Effectiveness of single-pass processing for long sequences.** We compare directly processing entire 270-frame sequences with our frequency-modulated position encoding (Single-Pass) against traditional sliding window approach with overlapping frames.

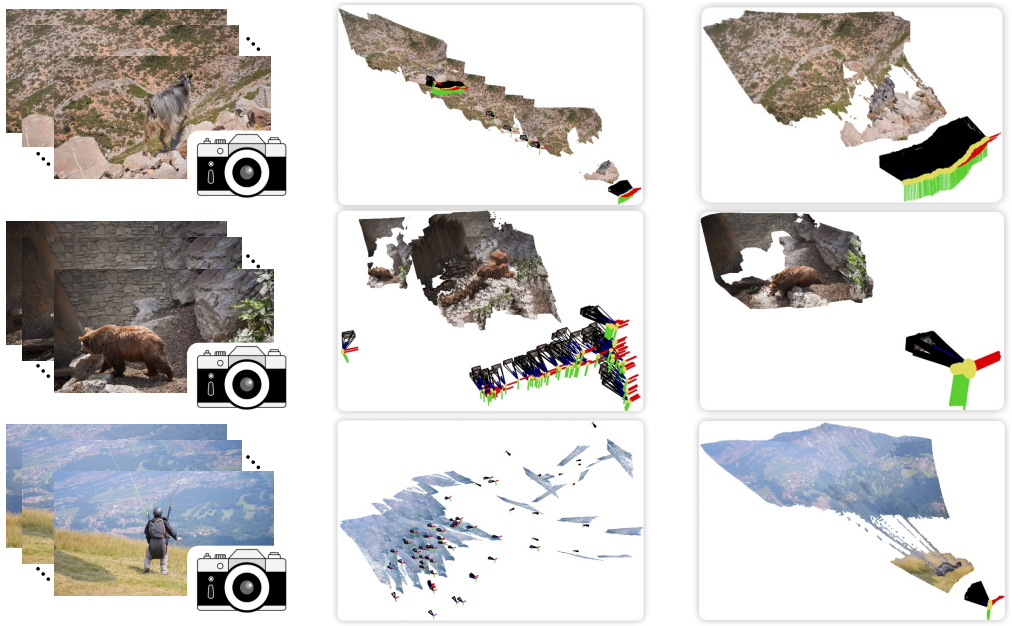

Figure 4: **4D reconstruction comparison using MegaSAM (Li et al., 2025).** Our method enables coherent multi-view reconstruction from video sequences (right), while MoGe (middle) produces fragmented results with significant distortion. Input video frames shown on left.

Quantitative results. Table 1 presents our method's performance compared to state-of-the-art approaches across diverse datasets. We report both point map estimation (Rel^p, δ^p) and depth estimation (Rel^d, δ^d) metrics, where Rel measures relative absolute error (lower is better) and δ represents the percentage of pixels with relative error less than 1.25 (higher is better). For temporal

consistency, we employ the temporal alignment error (TAE) metric introduced by Yang et al. (2025). All evaluations use a shared scale factor for alignment across entire video sequences to fairly assess global consistency.

Our approach significantly outperforms previous methods, achieving the lowest average rank across all datasets. Specifically, we achieve substantial improvements: 8.5% Rel^p reduction on Sintel, 24.2% accuracy and 34.9% temporal consistency (TAE^p) improvements on ScanNet, and 9.9% Rel^p reduction on KITTI compared to MoGe. Depth metrics show similar trends across datasets.

Table 3 evaluates our method against state-of-the-art video depth estimation approaches. Our method achieves competitive or superior accuracy across all datasets while demonstrating remarkable computational efficiency, processing sequences 6-42× faster than existing methods.

Qualitative comparison. Figure 3 presents spacetime slice visualizations where our method maintains superior temporal consistency across diverse scenarios compared to VGGT and MoGe. Figure 4 demonstrates the downstream impact, with our approach enabling coherent 4D reconstructions via MegaSAM (Li et al., 2025) while MoGe produces fragmented results under identical conditions. To quantify this reconstruction quality, we evaluated camera pose accuracy using our predicted point maps as input to MegaSAM on the Sintel dataset. Our method achieves significant improvements with ATE of 0.035 and RTE of 0.014, outperforming both MonST3R (ATE: 0.078, RTE: 0.038) and MoGe (ATE: 0.087, RTE: 0.033) by 55% and 60% respectively in ATE, and 63% and 58% respectively in RTE. Notably, our method achieved 100% success rate while MoGe failed completely on 2 scenes.

4.3 Ablation Study

Extrapolation strategies for long sequences. Table 2 analyzes different position encoding strategies for processing sequences significantly longer than our 24-frame training examples. We evaluate four approaches: no temporal position encoding (None), absolute position encoding (APE), standard rotary position encoding with NTK adaptation (RoPE), and our complete approach that combines NTK-adapted RoPE with sequence stretching during training to simulate extrapolation (RoPE+).

Effectiveness of temporal and geometric constraints. We evaluate our hierarchical temporal supervision ($\mathcal{L}_{temp}$) and cross-frame geometric constraints ($\mathcal{L}_{cross}$) on Sintel, ScanNet, and DDAD datasets. The combination of both losses achieves pointmap temporal consistency improvements of 9.53% on average across datasets and depth temporal consistency improvements of 18.4% compared to baseline MoGe constraints.

Single-pass vs. sliding window inference. Table 4 compares our single-pass processing approach with traditional sliding window techniques (Chen et al., 2025) for handling long sequences. Our method directly processes entire 270-frame sequences in a single forward pass. This approach not only eliminates computational redundancy but also consistently improves performance across all datasets. On KITTI, single-pass processing reduces Rel^p by 9.8% compared to sliding window approaches, highlighting the benefits of maintaining global context across the entire sequence rather than processing overlapping segments independently.

Computational efficiency. Using an NVIDIA H20 GPU with FP16 inference, our model processes 300 frames at 378×672 resolution in 7.68 seconds (39.1 FPS) with 76.53 GB memory usage. The optimization for 300 frames uses 0.337 seconds, averaging 1.12 ms per frame.

5 Conclusion

We presented a novel approach for monocular video geometry estimation that addresses the dual challenge of high geometric accuracy and long-range temporal consistency. Our method generates scale-invariant 3D point maps through three key innovations: viewpoint-invariant geometry aligning points in a unified reference frame, appearance-invariant learning preserving structural features despite visual variations, and frequency-modulated positioning enabling extrapolation to sequences vastly exceeding training examples. Experiments demonstrate substantial improvements over state-of-the-art methods, with our efficient single-pass approach maintaining both fine-grained detail and global consistency across diverse datasets. **Limitation**: Our current approach relies on external methods (Li et al., 2025) for camera pose estimation rather than direct prediction within our model, which will be addressed in future work toward a fully end-to-end solution.

REFERENCES

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, 2023.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *ICLR*, 2025.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv:2001.10773*, 2020.

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv:2306.15595*, 2023.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *CVPRW*, 2019.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

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. *arXiv:2403.12013*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv:2403.13788*, 2024.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *TPAMI*, 2024.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, 2018.

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.

Zhengfei Kuang, Tianyuan Zhang, Kai Zhang, Hao Tan, Sai Bi, Yiwei Hu, Zexiang Xu, Milos Hasan, Gordon Wetzstein, and Fujun Luan. Buffer anytime: Zero-shot video depth and normal from image priors. In *CVPR*, 2025.

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 *CVPR*, 2025.

LightwheelAI and LightwheelOcc contributors. Lightwheelocc: A 3d occupancy synthetic dataset in autonomous driving. `https://github.com/OpenDriveLab/LightwheelOcc`, 2024.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *CVPR*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023.

E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *IROS*, 2019.

Bowen Peng and Jeffrey Quesnelle. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *ICLR*, 2024.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv:2502.20110*, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2022.

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 *ICCV*, 2021.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. In *CVPR*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv:2104.09864*, 2021.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv:2212.10554*, 2022.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *CVPR*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *arXiv:1909.05452*, 2019.

Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *ICME*, 2021.

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*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024a.

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 *IROS*, 2020.

Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *ICCV*, 2023.

Yiran Wang, Min Shi, Jiaqi Li, Chaoyi Hong, Zihao Huang, Juewen Peng, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Nvds^{+}: Towards efficient and versatile neural stabilizer for video depth estimation. *TPAMI*, 2024b.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. In *ICLR*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024b.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *CVPR*, 2023.

Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019.

Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *TPAMI*, 2021a.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021b.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023.

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.

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 *ICCV*, 2023.