This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Stabilizing Streaming Video Geometry via Dynamic Feature Normalization

Xiaoyang Lyu*[1] Muxin Liu*[1] Xiaoshan Wu[1] Ruicheng Wang[2]
Yi-Hua Huang[1] Yang-Tian Sun[1] Shaoshuai Shi[3] Xiaojuan Qi[1◇]

[1]The University of Hong Kong [2]USTC [3]Voyager Research, Didi Chuxing

* Equal Contribution: {shawlyu, mxliu}@connect.hku.hk ◇ Corresponding Author: xjqi@eee.hku.hk

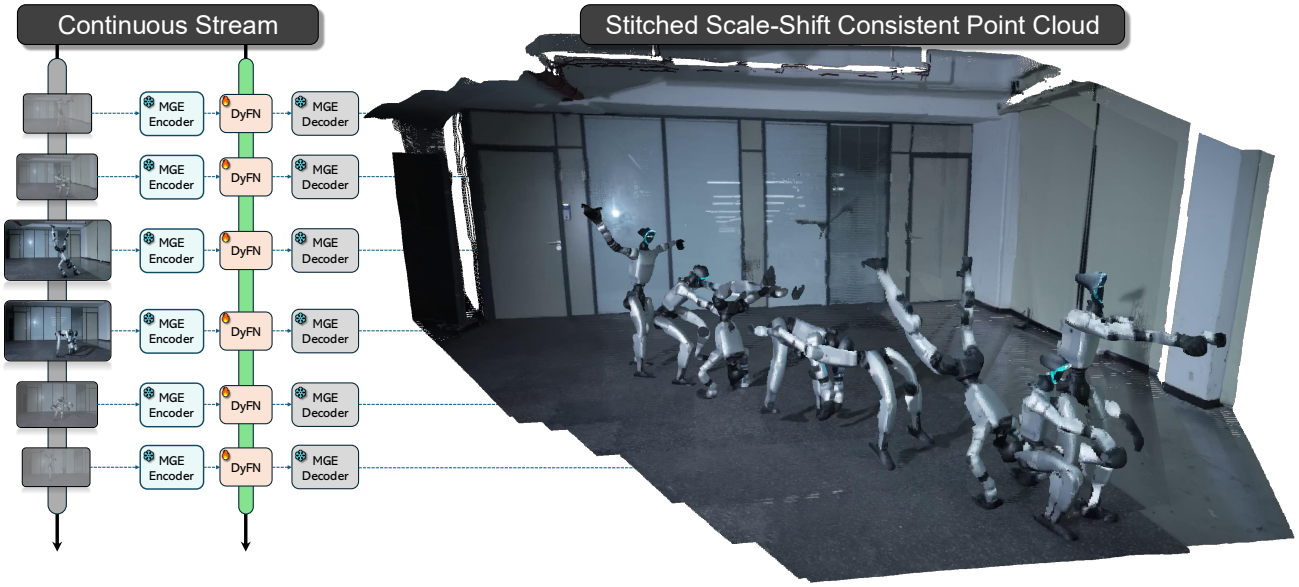


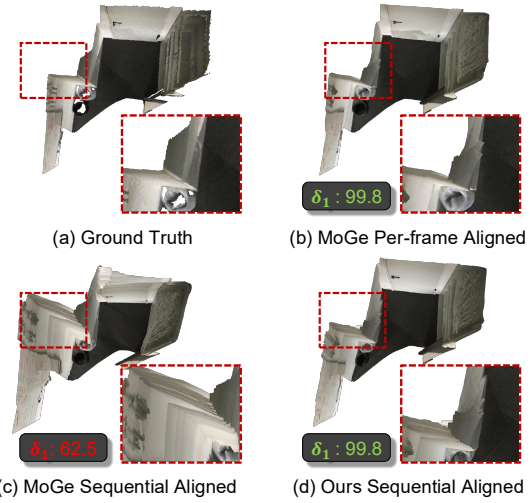
Figure 1. We introduce Dynamic Feature Normalization (DyFN), a novel module that transforms single-image geometry estimators into consistent streaming video models. DyFN achieves scale-shift consistency across continuous frames, enabling the generation of coherent 3D reconstructions from streaming monocular input.

Abstract

Consistent 3D geometry estimation from streaming RGB input is crucial for real-world applications such as autonomous driving, embodied AI, and large-scale reconstruction. While modern monocular geometry foundation models achieve strong single-image accuracy, they exhibit severe temporal inconsistency on continuous input, notably dominated by scale–shift drifting. Through targeted empirical analysis, we trace this instability to its root cause: fluctuations in latent feature statistics, whose mean and variance directly determine the predicted depth's scale and shift. Building on this insight, we introduce Dynamic Feature Normalization (DyFN), a lightweight, causal recurrent module that dynamically and robustly modulates feature statistics to maintain stable geometry over time. We adapt powerful pretrained monocular geometry models for streaming by finetuning only DyFN, a mere 2% additional parameters, while keeping the backbone frozen, thereby achieving temporal consistency without compromising single-image accuracy. Extensive experiments across four benchmarks show that DyFN effectively eliminates temporal artifacts such as disjointed layering and positional jitter, and achieves state-of-the-art temporal stability, improving over prior streaming methods by up to 14% and even outperforming heavier non-causal video baselines. Project page: https://shawlyu.github.io/DyFN

1. Introduction

3D geometry estimation is fundamental to many real-world applications, such as robotics, autonomous driving, and augmented reality. Recently, Monocular Geometry Estimation (MGE) and Monocular Depth Estimation (MDE) [2, 32, 35, 46, 47, 54, 55, 57] has progressed rapidly with the rise of large-scale foundation models, significantly narrowing the gap between single-image prediction and sensor-



(a) Ground Truth (b) MoGe Per-frame Aligned (c) MoGe Sequential Aligned (d) Ours Sequential Aligned

Figure 2. **Reconstruction comparison.** We align the predicted depth to metric scale using an affine transformation. Per-frame Aligned involves calculating the scale and shift for each frame independently. Sequence-aligned involves calculating a single, consistent scale and shift for the entire sequence. The point clouds are then fused using ground truth poses.(δ_1 means $\delta < 1.25$)

based measurements. Models such as MoGe [46] further exhibit remarkable zero-shot generalization across diverse scenes by learning geometry-aware priors from massive image collections. Despite these advances, most MGE and MDE models are designed for image inference, limiting their applicability in dynamic environments where input naturally arrives as continuous video streams.

When applied to continuous video streams, image-based MGE models exhibit pronounced temporal inconsistency: geometry predictions fluctuate across frames, causing distortions such as layering breaks and positional jitter in reconstructed scenes (Fig. 2c). Existing methods mitigate this issue using temporal attention [22, 38] or recurrent memory modules [11, 45] to enforce inter-frame coherence. However, these solutions come with notable limitations: they typically require full-network finetuning on large-scale annotated video datasets, which is computationally expensive and data-intensive; and such finetuning often degrades the per-frame accuracy and zero-shot generalization of pretrained MGE models by overfitting the backbone to specific video domains.

In this paper, we argue that existing foundation MGE models are already well-equipped to be repurposed for streaming video depth estimation without retraining the entire network. We posit that temporal inconsistency does not primarily arise from per-frame perceptual errors, but from inherent scale ambiguity: without a mechanism to maintain consistent scale and shift over time, the model effectively predicts each frame in an independently drifting coordinate system. Notably, pretrained models such as MoGe [46] already encode strong geometric structure—when each frame is *individually aligned with a simple per-frame scale and shift*, the reconstructed 3D geometry becomes accurate and temporally coherent (Fig. 2b). This implies that temporal inconsistency is largely driven by *frame-to-frame scale–shift instability* rather than deficient geometry. To further investigate this phenomenon, we conduct an empirical analysis of how global latent feature statistics influence predicted scale and shift (Sec. 3). Our results reveal that *scale/shift variations are tightly coupled with the mean and variance of latent features* extracted by the pretrained encoder (Fig. 3). This finding suggests that rather than retraining the entire model, temporal stability can be achieved by directly regulating these latent feature statistics of MGE models, specifically, their mean and variance.

Motivated by this observation, we propose **Dynamic Feature Normalization** (**DyFN**), a lightweight, learnable module that predicts and dynamically modulates the mean and variance of latent features over time to enforce consistent depth across frames, without compromising the fidelity or generalization of pretrained MGE models. For *online streaming depth estimation*, DyFN incorporates a recurrent ConvGRU-based module that updates normalization parameters based on aggregated historical context. By finetuning only this **DyFN**, while keeping the pretrained encoder and decoder frozen, our method efficiently adapts existing MGE models to continuous inputs, achieving temporal coherence without sacrificing geometric accuracy. Extensive experiments across diverse benchmarks demonstrate that **DyFN** achieves state-of-the-art temporal stability in streaming scenarios (See Figure 5 and Table 1). It not only surpasses strong video-based methods, but also preserves the strong single-frame accuracy of pretrained MGE models.

In summary, our main contributions are threefold:

- We identify the principal source of temporal instability in pretrained monocular depth models: *frame-to-frame scale–shift inconsistency* arising from fluctuations in latent feature statistics (mean and variance), thereby establishing a direct connection between feature distribution and temporal stability.
- We propose **Dynamic Feature Normalization (DyFN)**, a lightweight and general stabilization module that dynamically modulates latent feature statistics to maintain consistent scale and shift over time.
- We show that finetuning only this small stabilizer while freezing the pretrained MGE backbone achieves state-of-the-art temporal stability across diverse video benchmarks *without degrading* single-frame accuracy or generalization, surpassing fully trained video-based models.

2. Related Work

Relative Depth Estimation Relative depth estimation has demonstrated robust generalization across diverse domains

by predicting depth up to an unknown scale and shift [3, 13, 26, 35, 36, 54, 55]. Foundational works like MiDaS [35] established this paradigm using multi-dataset training combined with scale-invariant losses. Subsequent research has increasingly integrated large-scale self-supervised pre-training [19, 30, 50–52] to enhance feature representation. Notably, DPT [36] successfully adapted transformers for dense prediction, a strategy further scaled by Depth Anything [54, 55] utilizing over 60 million unlabeled images to achieve superior zero-shot performance. More recently, diffusion-based methods such as Marigold [25] and GeoWizard [15] have leveraged generative priors [34, 37] to push the boundaries of detail recovery. Despite these advancements, these frame-centric approaches process images independently, inherently failing to maintain temporal consistency when applied to video streams.

Metric Depth Estimation. To resolve scale-shift ambiguity, metric depth estimation aims to recover absolute depth from monocular inputs, a task that remains fundamentally ill-posed [2, 18, 21, 32, 33, 47, 56, 57]. Recent methods tackle this by incorporating strong geometric priors or optimizing for camera intrinsics. For instance, LeReS [56] leverages scene statistics to align predictions, while ZoeDepth [2] extends relative depth networks with adaptive metric bins to handle scene variability. Addressing the dependency on camera parameters, Metric3D [57] and its successor [21] propose zero-shot inference within a canonical camera space, whereas UniDepth [32] employs spherical parameterization to disentangle intrinsics for broader generalization. Although anchoring predictions to a metric scale theoretically reduces global scale drift, these methods process video frames individually. Without explicit temporal integration, they remain susceptible to inter-frame flickering and metric instability when applied to dynamic video streams.

Video and Stream Depth Estimation To explicitly model temporal dependencies, recent works extend image-based baselines by injecting bidirectional attention [40], incorporating recurrent networks [1, 17, 20, 45], or leveraging pre-trained video generative models [4, 41]. For instance, Video Depth Anything [8] augments the static Depth Anything V2 architecture with spatial-temporal attention and keyframe scheduling to enhance long-term consistency. Similarly, RollingDepth [24] employs multi-frame cross-attention aligned with global optimization. In the generative domain, ChronoDepth [38] pioneers the use of video diffusion priors for depth regression, while DepthCrafter [22] adopts a curriculum-based training strategy to synthesize temporally coherent sequences. However, these window-based methods typically rely on processing fixed-length clips with overlapping inference. This paradigm inherently incurs high latency and memory redundancy, limiting their applicability for long-duration or real-time scenarios. Consequently, streaming architectures have emerged as an efficient alternative, processing arbitrary sequence lengths via recurrent states. A representative work, FlashDepth [11], demonstrates this potential by maintaining a compact hidden state to achieve real-time inference at 2K resolution without sacrificing temporal stability.

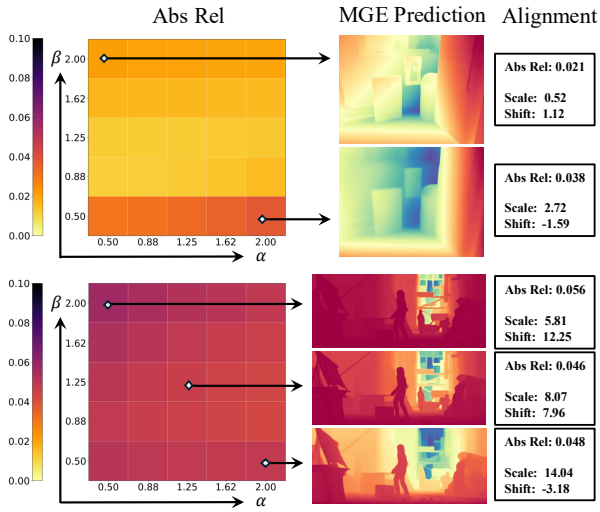


Figure 3. Empirical study on scale-shift variations. The left part illustrates the MGE performance (AbsRel) after modulating the latent features with sampled modulation parameters (α, β). The right part visualizes the corresponding MGE outputs and parameters used to align with GT. It can be observed that despite the predicted depth exhibiting significant scale and shift fluctuations, the underlying geometric accuracy remains largely unchanged.

Multi-frame Geometry Estimation Distinct from direct depth regression, geometric estimation methods reconstruct dense 3D point maps directly from images, leveraging explicit geometric constraints to enhance structural fidelity [7, 10, 28, 39, 42, 43, 45–48, 53, 58]. The foundational work, Dust3R [48], reformulates pairwise structure-from-motion as a regression task, enabling robust reconstruction from uncalibrated views. This paradigm has been extended to dynamic and sequential contexts: MonST3R [58] incorporates motion priors to handle dynamic objects, while VGGT [43] utilizes global spatiotemporal transformers for consistent scene reconstruction. To address the memory bottlenecks in long-sequence processing, CUT3R [45] adopts LSTM-style recurrent updates to accumulate geometric features, and TTT3R [10] optimizes memory read-write mechanisms to enhance localization stability. However, these geometry-centric approaches generally incur significant computational overhead and heavily rely on multi-view overlap. Consequently, they often struggle in pure monocular settings or highly dynamic environments where consistent geometric constraints are violated or absent.

3. Empirical Studies

Pretrained monocular geometry foundation models such as MoGe [46] achieve remarkable accuracy on single images, reconstructing fine geometric details and demonstrating strong zero-shot generalization. Despite this single-frame success, their direct application to continuous video streams reveals a critical limitation. When applied naively in a frame-by-frame manner, these models suffer from severe scale-shift ambiguity across consecutive frames. Although each individual prediction may appear geometrically plausible in isolation, they are not anchored to a consistent 3D coordinate frame. This failure to maintain a stable geometric reference results in significant structural instability in the aggregated 3D scene. As illustrated in Fig. 2c, this manifests as **non-rigid warping and geometric drift** over time, rather than a coherent, stable reconstruction. This discrepancy highlights a fundamental gap: while existing models encode powerful static spatial priors, they lack the cross-frame geometric coherence required for stable 3D reconstruction from continuous stream.

3.1. What are the causes of temporal inconsistency?

To understand the root cause of this inconsistency, we perform an empirical study using the state-of-the-art monocular depth estimator MoGe [46] as a representative model. We find that the model's underlying geometric understanding is already robust. When each frame's prediction is individually aligned to the ground truth using a simple affine transformation (scale and shift), the reconstructed 3D geometry becomes highly accurate and geometric consistent (Fig. 2b). As demonstrated in Fig 2c, when fusing the predicted point clouds using a single sequential alignment (one scale/shift for the entire sequence), the reconstruction suffers from severe non-rigid warping and geometric drift, achieving an accuracy ($\delta < 1.25$) of only 62.5. In stark contrast, when we align each frame individually (per-frame scale and shift), the accuracy dramatically increases to 99.8. This reveals that the primary cause of temporal inconsistency is not geometric degradation but rather *frame-to-frame scale-shift variation*, the predicted global depth scale and offset drift over time, leading to unstable geometry estimation sequences.

3.2. What are the causes of scale-shift variations?

We further investigate what causes such *scale–shift fluctuations* in monocular geometry models. Most modern MGE networks share a common encoder–decoder architecture, where an input image $\mathcal{I}$ is first encoded into latent features $\mathcal{F} = \mathcal{E}(\mathcal{I})$, which are then decoded into a point map $\mathcal{P} = \mathcal{D}(\mathcal{F})$.

Analysis Setup. As the scale and shift parameters are global statistics that govern the predictions, we study how the global *latent feature distribution* influences these parameters across frames. To examine this, we select two images from distinct domains (indoor and outdoor) and extract their latent features $\mathcal{F}$. For each feature map, we compute its channel-wise mean $\boldsymbol{\mu}_{\mathcal{F}}$ and standard deviation $\boldsymbol{\sigma}_{\mathcal{F}}$, then normalize the features as

$$\mathcal{F}_{\text{norm}} = \frac{\mathcal{F} - \boldsymbol{\mu}_{\mathcal{F}}}{\boldsymbol{\sigma}_{\mathcal{F}} + \epsilon}, \tag{1}$$

where ϵ is a small constant for numerical stability. We then introduce scaling multipliers $\alpha, \beta \in [0.5, 2.0]$ to modulate the mean and standard deviation, defining

$$\boldsymbol{\mu}_{\mathcal{F}}^{\alpha} = \alpha \cdot \boldsymbol{\mu}_{\mathcal{F}}, \qquad \boldsymbol{\sigma}_{\mathcal{F}}^{\beta} = \beta \cdot \boldsymbol{\sigma}_{\mathcal{F}}, \tag{2}$$

and reconstructing modified features as

$$\mathcal{F}^{\alpha,\beta} = \mathcal{F}_{\text{norm}} \cdot \boldsymbol{\sigma}_{\mathcal{F}}^{\beta} + \boldsymbol{\mu}_{\mathcal{F}}^{\alpha}. \tag{3}$$

Each modified feature map $\mathcal{F}^{\alpha,\beta}$ is fed through the frozen decoder $\mathcal{D}$ to produce a new point map. Using least-squares fitting, we estimate the corresponding affine transformation (scale and shift) relative to the original output. The correlation between the modulation parameters (α, β) and the resulting geometric transformation is visualized in Fig. 3.

Empirical Results. As shown in Fig. 3, our experiments on both indoor (ScanNet) and outdoor (Sintel) datasets show a clear phenomenon. We found that altering the statistics of latent features (such as their mean and variance) directly causes the absolute scale and shift of the predicted depth maps to change dramatically (ranging from [0.52, 14.04] and [-3.18, 12.25], respectively). However, even when the scale and shift changed this much, the actual geometric shape of the prediction remained surprisingly stable. We verified this by applying the affine alignment to match the predictions with the ground truth; after alignment, the geometric accuracy was still very high. These results confirm our central hypothesis: *the mean and variance of latent features are strongly coupled with the prediction's scale and shift*, but are *largely decoupled from its relative geometric accuracy*. In essence, uncontrolled fluctuations in these feature statistics across a video stream are the direct cause of the observed scale-shift drift, which in turn manifests as the structural inconsistency detailed previously. This key insight directly motivates our proposed *Dynamic Feature Normalization*, a lightweight mechanism designed to explicitly regulate these statistics over time to enforce stable and geometrically coherent point predictions, as detailed in the following section.

4. Dynamic Feature Normalization

Inspired by the empirical findings in Sec. 3, we propose the Dynamic Feature Normalization (DyFN) module to address

the scale-shift inconsistency inherent to pretrained monocular geometry model. DyFN dynamically modulates latent features based on their temporal context to enforce stable and consistent geometry predictions for *online streaming video*. As illustrated in Fig. 4, our approach freezes the pretrained encoder and decoder. Only the lightweight DyFN module is trained, allowing it to adapt the feature statistics for temporal consistency.

This module first normalizes the incoming latent feature $\mathcal{F}_t$ to obtain a standardized feature $\mathcal{F}_t^{\text{norm}}$ following the Eq. 1. This feature is fed into a Convolutional GRU (ConvGRU) [1], which maintains a hidden state $\boldsymbol{h}_t$ that summarizes observations from all previous frames. At each timestep t, the ConvGRU updates its state based on the previous state $\boldsymbol{h}_{t-1}$ and the current feature $\mathcal{F}_t$:

$$\boldsymbol{h}_t = \text{ConvGRU}(\mathcal{F}_t, \boldsymbol{h}_{t-1}). \tag{4}$$

For the first frame ($t = 1$), the hidden state $\boldsymbol{h}_0$ is initialized as a zero tensor.

Two lightweight 1×1 convolutional heads then project the hidden state $\boldsymbol{h}_t$ to predict the spatial modulation parameters, a mean $\hat{\boldsymbol{\mu}}_t$ and a standard deviation $\hat{\boldsymbol{\sigma}}_t$:

$$\hat{\boldsymbol{\sigma}}_t = \text{Conv}^{\sigma}_{1\times1}(\boldsymbol{h}_t), \quad \hat{\boldsymbol{\mu}}_t = \text{Conv}^{\mu}_{1\times1}(\boldsymbol{h}_t). \tag{5}$$

These predicted statistics are used to modulate the normalized feature, effectively replacing the original, unstable per-frame statistics with temporally-aware ones:

$$\mathcal{F}_t^{\text{consistent}} = \hat{\boldsymbol{\sigma}}_t \cdot \mathcal{F}_t^{\text{norm}} + \hat{\boldsymbol{\mu}}_t. \tag{6}$$

This final re-modulated feature $\mathcal{F}_t^{\text{consistent}}$ is then passed to the frozen decoder $\mathcal{D}$ to predict the final depth map P_t.

5. Model Training

Trained Modules. Guided by our analysis that the pretrained monocular geometry model already captures robust *relative* geometric knowledge, we adopt a parameter-efficient fine-tuning strategy. We freeze the weights of both the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ to preserve their powerful, generalized feature representations. Temporal consistency is then achieved by training **only** the lightweight DyFN module, which is tasked with modulating the latent feature distributions. This approach is highly efficient, as the DyFN module constitutes merely 2% of the total parameters. As demonstrated in Tab. 1 and Tab. 2, this strategy successfully achieves our dual objectives: it retains the strong per-frame geometric accuracy of the frozen backbone while efficiently enforcing sequence-level stability.

Training Objective. In addition to the base loss from the backbone, $\mathcal{L}_{\text{MoGe}}$, we introduce two terms that explicitly supervise scale-shift stability. The first is a *global alignment loss*, $\mathcal{L}_{\text{align}}$, designed to enforce a single, consistent scale and shift across the entire sequence. Given a sequence of L predicted point maps $\{\hat{\text{P}}_j\}_{j=1}^{L}$, we compute a *single* global affine pair (s_g, t_g) that best aligns all points from all frames to the ground truth simultaneously (e.g., via least-squares). The loss is then defined as the total error measured *only* against this single global transformation:

$$\mathcal{L}_{\text{align}} = \sum_{j=1}^{L} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \left\| s_g \hat{\text{p}}_j^{\,i} + t_g - \text{p}_{\text{j}}^{\text{i}} \right\|_1, \tag{7}$$

where $\text{p}_{\text{j}}^{\text{i}}$ is the i-th ground-truth point in frame j with depth z_i, $\hat{\text{p}}_j^{\,i}$ is the corresponding prediction, and $\mathcal{M}$ denotes valid points. This formulation directly penalizes any frame j whose prediction $\hat{\text{p}}_j$ deviates from the sequence-wide optimal scale s_g and shift t_g. This forces the network's underlying feature representation (e.g., via DyFN) to produce outputs that are inherently stable over time.

The second term is an *inter-frame temporal loss*, $\mathcal{L}_{\text{temp}}$, designed to mitigate long-horizon drift. It enforces that the **magnitude of temporal change** in the predictions matches that of the ground truth. To capture both short- and long-term dynamics, we compute this loss over multiple window sizes $k \in K = \{1, 2, 4\}$. The loss penalizes the scaled L1-discrepancy between the predicted and ground-truth inter-frame deltas:

$$\mathcal{L}_{\text{temp}} = \sum_{k \in K} \sum_{j=1}^{L-k} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \left\| s_g \hat{\delta}_{j,k}^{\,i} - \delta_{j,k}^{\,i} \right\|_1, \tag{8}$$

where $\hat{\delta}_{j,k}^{\,i} = \left\| \hat{\text{p}}_j^{\,i} - \hat{\text{p}}_{j+k}^{\,i} \right\|_1$ represents the magnitude of the predicted change for point i across k frames, and $\delta_{j,k}^{\,i} = \left\| \text{p}_{\text{j}}^{\text{i}} - \text{p}_{\text{j+k}}^{\text{i}} \right\|_1$ is the corresponding ground-truth change. Applying the global scale s_g (derived from $\mathcal{L}_{\text{align}}$) ensures that the predicted deltas are measured in the same metric space as the ground-truth deltas before comparison.

Our final objective is a weighted sum:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{MoGe}} + \alpha\, \mathcal{L}_{\text{align}} + \beta\, \mathcal{L}_{\text{temp}}, \tag{9}$$

where $\alpha = 1$ and $\beta = 0.1$ in our training. Details of $\mathcal{L}_{\text{MoGe}}$ are provided in the supplementary material.

Training Data. To ensure robustness across diverse scenarios, we finetune only the DyFN module on a large-scale data compilation. This combined corpus provides approximately 1M total frames for training. Unless otherwise noted, our training procedure involves sampling fixed-length, continuous clips of 12 frames. More details are shown in the supplementary material.

6. Experiment

Baselines To comprehensively evaluate our method on video depth estimation, we compare it against representa-

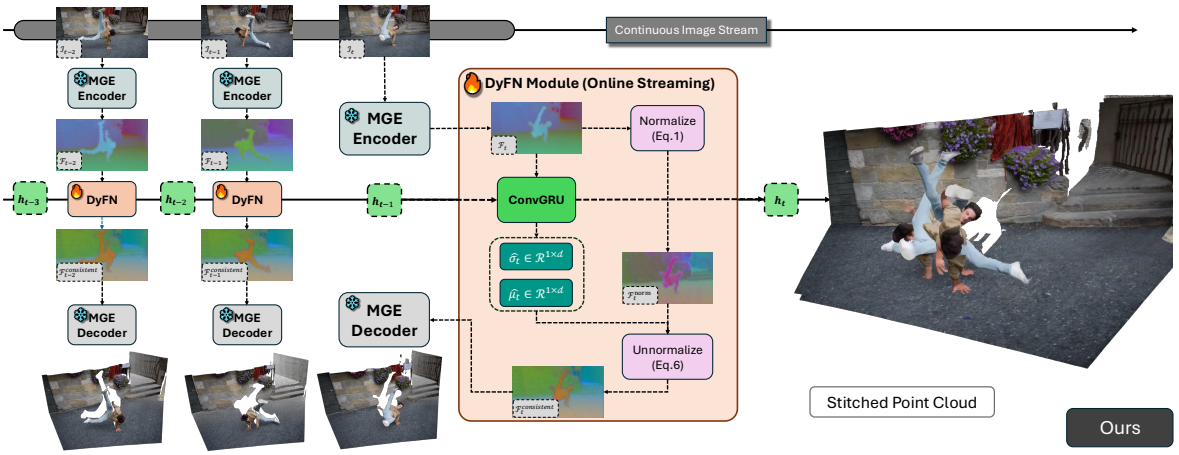


Figure 4. **Method Overview.** Our method performs consistent geometry estimation from an image stream. Each frame is processed by a shared ViT-based MGE encoder to extract visual features $\mathcal{F}_t$. These features are then passed into our recurrent *Dynamic Feature Normalization (DyFN)* module. The DyFN module leverages a temporally-aware hidden state $\boldsymbol{h}$ to dynamically modulate the mean and variance of $\mathcal{F}_t$, producing temporally consistent features $\mathcal{F}_t^{consistent}$. These stabilized features are subsequently fed into the MGE decoder to regress a consistent point map. Finally, a correspondence-based rigid pose solver (estimating rotation and translation) aggregates these point maps to produce a stable and coherent 3D reconstruction (See supplementary for more details).

tive works from six distinct paradigms. We broadly categorize these approaches based on their output: *Depth Estimation* methods produce only per-frame depth, while *Geometry* methods output per-frame point maps. The specific categories are as follows: (1) *Relative Depth Estimation:* Single-image models trained with an affine-invariant loss. As they process frames independently, they lack temporal consistency (e.g., Marigold [25], Depth Anything V1&V2 ("DAV1, DAV2") [54, 55], MoGe [46]). (2) *Metric Depth Estimation:* Single-image models that are trained to predict depth at a true metric scale (e.g., DepthPro [5], MoGe v2 [47]). (3) *Multi-frame Geometry:* Offline models that process a set of views and leverage mechanisms like bidirectional cross-attention to enforce multi-view consistency (e.g., VGGT [43], Monst3R [58]). (4) *Streaming Geometry:* Methods that support online inputs, fusing temporal information using recurrent modules(e.g., CUT3R [45], TTT3R [10]). (5) *Video Depth Estimation:* Models that use temporal fusion but are constrained to fixed-length, offline video inputs (e.g., DepthCrafter [22], VideoDepthAnything("VDA") [38]). (6) *Streaming Depth Estimation:* Methods support online streaming inputs and use recurrent modules to update features for temporal consistency (e.g., FlashDepth [11]).

Datasets and Metrics. For quantitative evaluation, we follow the protocol of DepthCrafter [22] and select representative scenes from datasets covering indoor [12, 31], outdoor [16], and in-the-wild environments [6]. We evaluate geometric accuracy using the Absolute Relative Error (AbsRel) and the $\delta < 1.25$ threshold. Detailed formulations for these metrics are available in the supplementary material.

Evaluation Protocol. For non-metric models, we first align their predictions to the ground-truth scale using a least-squares method. A crucial distinction lies in the scope of this alignment, which we define for two separate evaluation settings: **(1) Video Depth Evaluation:** We compute a single scale and shift for the entire sequence and apply it uniformly to all frames. This protocol stringently evaluates both per-frame accuracy and, critically, inter-frame temporal consistency. **(2) Image Depth Evaluation:** We compute a separate scale and shift for each frame independently. This protocol isolates per-frame accuracy, measuring the model's static performance without penalizing temporal instability.

Method	Sintel (50 frames)		Scannet (90 frames)		KITTI (110 frames)		Bonn (110 frames)	
	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
• Marigold	0.532	51.5	0.166	76.9	0.149	79.6	0.091	93.1
• DAV1	0.325	56.4	0.130	83.8	0.142	80.3	0.078	93.9
• DAV2	0.367	55.4	0.135	82.2	0.140	80.4	0.106	92.1
• MoGe v1	0.216	65.3	0.117	84.7	0.076	96.0	0.074	95.5
• DepthPro	0.319	52.0	(0.088)	(92.7)	(0.088)	(92.2)	(0.063)	(96.6)
• MoGe v2	0.214	69.5	(0.110)	(88.2)	(0.183)	(58.8)	(0.049)	(98.0)
• VGGT	0.287	66.1	0.031	98.5	0.070	96.5	0.055	97.1
• Monst3R	0.335	58.5	0.123	83.2	0.104	89.5	0.063	96.4
• CUT3R	0.421	47.9	0.097	88.7	0.118	88.1	0.078	93.7
• TTT3R	0.404	50.0	0.114	87.7	0.113	90.4	0.068	95.4
• DepthCrafter	0.270	69.7	0.123	85.6	0.104	89.6	0.071	97.2
• VDA	0.300	63.3	0.075	95.4	0.079	95.0	0.051	98.1
• FlashDepth	0.265	64.2	0.101	90.3	0.103	89.5	0.053	98.0
• Ours	**0.180**	**73.0**	**0.073**	**96.6**	**0.062**	**97.3**	**0.044**	**98.4**

Table 1. **Quantitative evaluation of video depth estimation** on the Sintel, ScanNet, KITTI, and Bonn datasets. We compare methods across six categories: • Relative Depth, • Metric Depth, • Multi-frame Geometry, • Streaming Geometry, • Video Depth, and • Streaming Depth. The best results are highlighted in **bold**. Values in gray indicate that the method was trained on the target dataset. Values in parentheses denote evaluations performed on the raw metric output without alignment.

6.1. Monocular and Video Depth Estimation

Video Depth Estimation As shown in Table 1, we conduct a comprehensive quantitative comparison of our proposed method against six distinct categories of existing works across four benchmarks (Sintel, ScanNet, KITTI, and Bonn). The results clearly demonstrate that our method achieves state-of-the-art performance, outperforming all other models across all datasets and metrics (AbsRel ↓ and $\delta < 1.25$ ↑). (1) Comparison with Monocular Depth Models (Categories • and •): Single-image models suffer from a lack of temporal constraints. Their inherent scale-shift inconsistency is heavily penalized by the video evaluation protocol, which uses a single scale and shift for the entire sequence. Our method, by explicitly enforcing temporal consistency, shows significant improvements; for example, on Scannet, our $\delta < 1.25$ (96.6) is a **11.9%** improvement over MoGe v1 (84.7). Metric depth models, while trained to align with a metric scale, are similarly unstable and lack temporal fusion. This leads to volatile performance across datasets. For instance, MoGe v2's AbsRel on KITTI is 0.183, drastically worse than our 0.062. This highlights the superior stability and consistency of our approach. (2) Comparison with Multi-view Geometry Estimator (Categories • and •): Multi-view geometry estimation methods mainly rely on static scene assumptions and known poses. Consequently, they perform poorly in dynamic scenes like Sintel (e.g., VGGT AbsRel 0.287 vs. our 0.180 on Sintel). Notably, while VGGT performs well on the static ScanNet dataset (0.031), this result is attributable to it being trained on this specific dataset (indicated by gray text) and is not indicative of generalizability. On the static Bonn benchmark, our method (0.044) still surpasses VGGT (0.055) on Abs Rel. Furthermore, streaming-based reconstruction methods show an even more significant performance drop, confirming the limitations of their temporal fusion mechanisms. (3) Comparison with Video/Streaming Depth Models (Categories • and •): As our empirical study suggests, our method addresses the root cause of temporal inconsistency through dynamic feature normalization. As a result, our performance is not only significantly better than other streaming-based methods like FlashDepth (e.g., 96.6 vs. 90.3 AbsRel on Scannet), but it also surpasses offline video depth methods that utilize more complex bidirectional attention mechanisms. The qualitative results in Fig. 5 provide visual confirmation. For this visualization, we align predictions to the metric scale and transform point clouds into the global coordinate system using ground-truth poses. Our method's reconstructions exhibit superior geometric coherence and markedly less non-rigid warping compared to both FlashDepth and Video Depth Anything. Consequently, our method sets a new state-of-the-art, demonstrating that regulating feature statistics, as motivated by our empirical study (Sec. 3), is a highly effective strategy for achieving robust 3D consistency from streaming input.

	Sintel		Scannet		KITTI		Bonn	
Method	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
• DAV2	0.200	74.1	0.039	98.2	0.073	95.3	0.048	98.0
• MoGe v1	**0.124**	**83.7**	**0.027**	**98.6**	**0.044**	**98.0**	**0.028**	**98.8**
• CUT3R	0.428	55.4	0.064	93.7	0.092	91.3	0.063	96.2
• VDA	0.200	75.3	0.041	98.1	0.074	95.1	0.039	98.6
• FlashDepth	0.174	75.6	0.056	96.3	0.085	92.6	0.043	98.7
• Ours	**0.124**	**83.7**	**0.027**	**98.6**	**0.044**	**98.0**	**0.028**	**98.8**

Table 2. Single-frame depth evaluation. We report performance across four different categories: • Relative Depth, • Streaming Geometry, • Video Depth, and • Streaming Depth. Evaluations are performed on the Sintel, ScanNet, KITTI, and Bonn datasets. All models accept only a single image as input at a time.

Single Frame Depth Estimation. In addition to video-based metrics, we conduct a single-frame depth evaluation, with results shown in Table 2. A key advantage of our methodology is that by freezing the pretrained encoder and decoder, our model **perfectly inherits the per-frame accuracy of its base model (MoGe v1)**. As the table demonstrates, our results are identical to MoGe v1 across all four datasets. This is a critical distinction from other finetuning approaches, which often suffer from accuracy degradation. For example, FlashDepth (which builds on DepthAnything v2) sees its $\delta < 1.25$ score on KITTI drop from 95.3 (the base model's score) to 92.6 after finetuning. Our method avoids this tradeoff entirely. By preserving the base model's state-of-the-art accuracy, our approach achieves the best results across all datasets when compared to all other video and streaming-based methods.

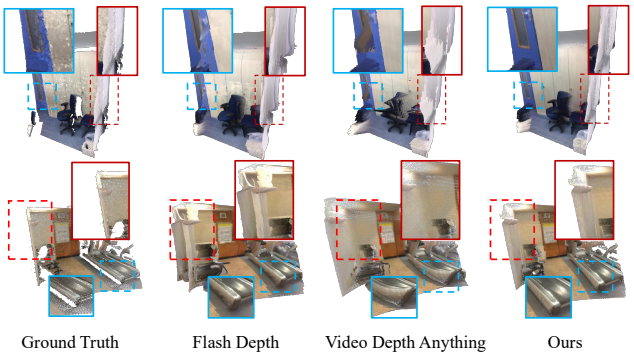


Figure 5. Qualitative comparison on indoor scenes. We compare our method with Flash Depth and VDA. Our method shows the best geometric consistent and less non-rigid warping.

6.2. Ablation Study

We conduct comprehensive ablation studies to validate our design choices, focusing on four key aspects: the effectiveness of the DyFN module, the contribution of loss functions, the choice of recurrent unit, and the global alignment strategy. Quantitative results are reported in Table 3. All ablated experiments are trained on the full dataset as described on Sec. 5, more details can be seen in the supplementary.

Effectiveness of DyFN. We demonstrate the impact of our proposed DyFN module. Compared to the MoGe, the in-

Method	Sintel		Scannet		KITTI		Bonn	
	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
•MoGe	0.216	65.3	0.117	84.7	0.076	96.0	0.074	95.5
•Ours†	0.180	73.0	0.073	96.6	0.062	97.3	0.044	98.4
•w/o $\mathcal{L}_{align}$	0.245	61.8	0.124	83.1	0.088	93.5	0.088	93.3
•w/o $\mathcal{L}_{temp}$	0.183	72.7	0.069	96.4	0.063	97.0	0.044	98.4
•DyFN (convgru) †	0.180	73.0	0.073	96.6	0.062	97.3	0.044	98.4
•DyFN (gru) †	0.187	72.5	0.078	94.9	0.065	96.8	0.053	98.1
•$\mathcal{L}^{\dagger}_{align}$	0.180	73.0	0.073	96.6	0.062	97.3	0.044	98.4
•$\mathcal{L}^{\ddagger}_{align}$	0.189	72.1	0.066	96.4	0.070	96.2	0.045	98.3

Table 3. Ablation studies. We conduct four different ablation studies, including: • Effectiveness of DyFN, • Impact of loss function, • Recurrent Unit Selection and • Global Scale Alignment Strategy. † means that the model was trained with first frame alignment strategy. ‡ means that the model was trained with the global frame alignment strategy.

troduction of DyFN yields significant improvements across all benchmarks. This confirms that the module effectively enhances temporal accuracy in video depth estimation.

Impact of Loss Functions. The removal of the alignment loss results in a sharp performance drop, even degrading accuracy beyond the baseline. This indicates that global-level alignment supervision is critical; without it, the DyFN module fails to learn the correct scale needed for alignment, negatively affecting relative depth precision. Additionally, the temporal loss ($\mathcal{L}_{\text{temp}}$) further refines the results by enforcing slight improvements in temporal consistency.

Recurrent Unit Selection. We compare different recurrent structures within DyFN. While both GRU and ConvGRU improve temporal consistency, the ConvGRU variant achieves superior performance. We attribute this to ConvGRU's ability to better capture spatially structured temporal statistics (i.e., stable mean and variance) compared to the standard GRU. See Supplementary for detailed results.

Global Scale Alignment Strategy. Finally, we evaluate the calculation method for the global scale s_g and shift t_g used in $\mathcal{L}_{align}$. We compare "First Frame Alignment" (denoted by †, aligning based on the first frame's prediction and GT) against "Global Alignment" (denoted by ‡, derived from the entire sequence). Results show that the First Frame strategy outperforms the Global strategy. We observe that using the full sequence for alignment complicates optimization, as initial predictions are unstable, and significantly increases training overhead. Consequently, we adopt the First Frame Alignment strategy for our final model.

6.3. Long Sequence Performance

To further demonstrate the robustness of our method against scale drift, we conducted experiments on 100 selected scenes from the ScanNet dataset [12], each comprising a continuous sequence of 500 frames. We compared our approach against two state-of-the-art baselines: FlashDepth [11] and VideoDepthAnything [9]. The evaluation was performed at incremental intervals of 100 frames. Crucially, to rigorously test global consistency, we recalculated the sequence-wise scale and shift alignment at each evaluation step. As illustrated in Fig. 6, while extended sequence lengths generally introduce accumulated error, our method demonstrates significantly superior stability. The *Ours* curve exhibits a much slower rate of degradation in both Absolute Relative Error (AbsRel) and Accuracy ($\delta < 1.25$) compared to the baselines. Notably, even at the 500-frame mark, our method preserves high accuracy and outperforms both FlashDepth and VDA by a clear margin. This empirically validates our method's effectiveness in maintaining scale consistency over long durations.

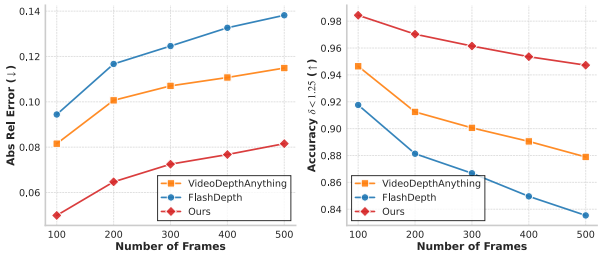


Figure 6. **Long-sequence robustness analysis.** We report the Abs Rel error (↓) and Accuracy ($\delta < 1.25, \uparrow$) evaluated at increasing frame intervals (100 to 500). Our method (red) demonstrates minimal performance decay compared to FlashDepth and VideoDepthAnything, maintaining superior scale consistency even as the sequence length increases.

7. Conclusion

We identified that temporal inconsistency in monocular geometry models stems from latent feature fluctuations causing scale-shift drift. We introduced Dynamic Feature Normalization, a lightweight recurrent module that stabilizes these feature statistics over time. By finetuning only DyFN while freezing the pretrained backbone, our method preserves single-frame accuracy while achieving state-of-the-art temporal stability. Our results demonstrate superior performance, even on long-duration sequences where we mitigate the error accumulation that plagues other methods. This work offers a simple and efficient path for adapting static foundation models to continuous video streams.

8. Acknowledgment

This work was conducted during an internship at Voyager Research, DiDi Chuxing. The research was supported by the Hong Kong Research Grants Council (RGC) through the General Research Fund (Grants No. 17202422, 17212923, and 17215025), the Theme-based Research Scheme (Grant No. T45-701/22-R), and the Strategic Topics Grant (Grant No. STG3/E-605/25-N). Additionally, part of this research was conducted at the JC STEM Lab of Robotics for Soft Materials, funded by The Hong Kong Jockey Club Charities Trust. The authors would also like to thank Yikang Ding and Xin Kong for their valuable advice and insightful discussions throughout the course of this work.

References

[1] Nicolas Ballas, Yao Li, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations (ICLR)*, 2016. 3, 5

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 3

[3] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 3, 4

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[5] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625, 2012. 6, 2

[7] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1050–1060, 2025. 3

[8] 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. *arXiv preprint arXiv:2501.12375*, 2025. 3

[9] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 8

[10] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 3, 6

[11] Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. *arXiv preprint arXiv:2504.07093*, 2025. 2, 3, 6, 8

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6, 8, 2

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

[14] 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. 2

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

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6, 2

[17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 3

[18] Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26996–27006, 2025. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 3

[21] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 3

[22] 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. *arXiv preprint arXiv:2409.02095*, 2024. 2, 3, 6

[23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023. 2

[24] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. *arXiv preprint arXiv:2411.19189*, 2024. 3

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3, 6

[26] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 3

[27] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[28] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 3

[29] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019. 2

[30] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[31] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2019. 6, 2

[32] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 1, 3

[33] 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, 2025. 3

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1, 3

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[38] 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, 2024. 2, 3, 6

[39] Yang-Tian Sun, Xin Yu, Zehuan Huang, Yi-Hua Huang, Yuan-Chen Guo, Ziyi Yang, Yan-Pei Cao, and Xiaojuan Qi. Unigeo: Taming video diffusion for unified consistent geometry estimation. *arXiv preprint arXiv:2505.24521*, 2025. 3

[40] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18782, 2023. 3

[41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[42] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3

[43] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 3, 6

[44] 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, 2021. 2

[45] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2, 3, 6

[46] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 1, 2, 4, 6

[47] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 1, 3, 6

[48] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[49] 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. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2

[50] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 3

[51] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela

Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023.

[52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[53] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 3

[54] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1, 3, 6

[55] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 3, 6

[56] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 3

[57] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 1, 3

[58] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3, 6

[59] 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. 2