GEOMETRY-GROUNDED VIDEO DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Our goal is to infer temporally and geometrically consistent depth maps that lift monocular video clips into 3D. To this end, we develop **G2**, a geometry-grounded feed-forward framework that fills in three missing pieces in video depth estimation, *i.e.*, cross-frame modelling, geometric supervision, and data scaling. First, instead of most existing efforts relying on fixed-location feature aggregation or externally estimated correspondences, we propose *Motion-Adaptive Implicit Alignment* to adaptively locate a sparse set of query-relevant features for efficient cross-frame information aggregation under camera and object motion. Second, we introduce *Explicit 3D Supervision* to directly constrain intra-frame structure and inter-frame geometry after back-projection of depth maps. Third, to scale geometric supervision to unlabeled videos, we develop *Multi-View Geometric Distillation*, which distills geometric knowledge from a reconstruction teacher to provide pseudo 3D supervision, yielding a more scalable solution. Extensive experiments on standard benchmarks demonstrate that G2 achieves state-of-the-art performance on both zero-shot video depth estimation and geometric reconstruction. We will release the code and model.

1 INTRODUCTION

Video depth estimation lifts monocular video clips to 3D by inferring dense depth at every frame. It plays a fundamental role in 3D perception, and has a broad range of applications, including robotics, drone technology, or augmented reality. In recent years, scaling model capacity and training data has driven substantial progress in single-image depth estimation, leading to the flourishing of depth foundation models (Yang et al., 2024a;b; Ke et al., 2025b). However, extending these advances to video remains challenging, as reliable video depth estimation requires not only spatially accurate, but also temporal consistency and geometrically faithful 3D structure across frames.

Recent video depth estimation methods explore dedicated mechanisms to improve the spatial accuracy and temporal consistency of depth predictions across adjacent frames. Existing solutions are typically built upon a *test-time optimization* scheme (*i.e.*, rely on camera poses or optical flow and optimize depth for each new video during inference) (Chen et al., 2019; Kopf et al., 2021; Luo et al., 2020; Zhang et al., 2021), adopt a *feed-forward prediction* paradigm (*i.e.*, directly estimate depth sequences from input videos) (Khan et al., 2023; Teed & Deng, 2018; Wang et al., 2022), and/or *repurpose generative video models* for depth prediction (Shao et al., 2025; Hu et al., 2025; Yang et al., 2025).

Unfortunately, these successful approaches do not guarantee geometrically faithful 3D structure, since they focus largely on view-dependent 2.5D representation while neglecting the cross-view alignment of the underlying 3D surfaces. Consequently, individually accurate depth maps may still yield spatially misaligned geometry when back-projected into 3D. This limitation largely stems from insufficient capabilities in three aspects: **cross-frame modelling**, **geometric supervision**, and **data scaling**. First, existing cross-frame modelling strategies (Chen et al., 2025; Chou et al., 2025; Sobko et al., 2026; Liu et al., 2026) often aggregate features at predefined locations, and struggle to establish reliable correspondence under camera or object motion. Explicit correspondences obtained from optical flow (Wang et al., 2025c; Han et al., 2026) or point tracks (Lai & Vedaldi, 2025; Cho et al., 2025) offer a direct solution, but require additional motion estimation or tracking. Second, existing supervision remains largely depth-centric, providing only indirect constraints on 3D geometry. Although some methods back-project predicted depth into 3D (Zhao et al., 2024; Yuan et al., 2025), they primarily use the resulting geometry for prediction consistency rather than explicit alignment with reference 3D structure. Third, existing methods face a fundamental data scaling bottleneck:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

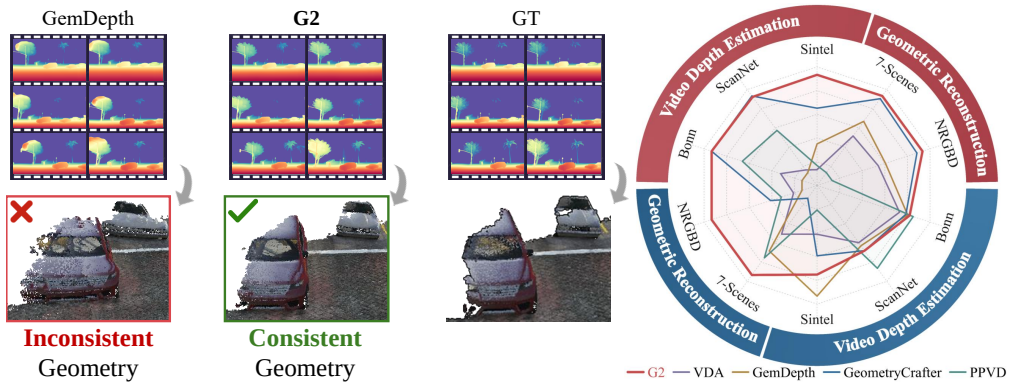


Figure 1: Our proposed **G2** produces temporally and geometrically consistent depth estimates, substantially outperforming existing state-of-the-art methods across 10 zero-shot video depth estimation and geometric reconstruction benchmarks.

high-quality video-depth pairs with accurate geometric supervision are costly to obtain at scale, limiting the benefits of data scaling that have driven recent advances in image depth estimation.

To fill in these three missing pieces to the puzzle of video depth estimation, we present a geometry-grounded feed-forward framework – **G2**. **First**, we introduce *motion-adaptive implicit alignment* (MAIA), which uses query-conditioned sparse sampling in the spatio-temporal feature volume to dynamically locate motion-displaced scene content. By implicitly establishing cross-frame feature correspondences, MAIA enables more efficient and reliable cross-frame information aggregation, improving depth estimation under camera and object motion. **Second**, we introduce *explicit 3D supervision* (E3DS), which back-projects predicted depths into 3D and directly anchors the recovered geometry to reference scene structure, by imposing an intra-frame point-map loss with an inter-frame projection loss. **Third**, to scale geometric supervision to unlabeled videos, we propose *multi-view geometric distillation* (MVGD), which distills geometric knowledge from a multi-view reconstruction teacher to provide scalable 3D supervision, enabling depth models to learn geometrically faithful representations from unlabeled videos.

These three contributions together make G2 a powerful framework to address temporally and geometrically consistent depth predictions. We evaluate it on the tasks of zero-shot video depth estimation and geometric reconstruction, and our approach achieves state-of-the-arts (see Fig. 1).

2 Related Work

Monocular depth estimation has progressed from dataset-specific prediction (Eigen et al., 2014; Eigen & Fergus, 2015; Fu et al., 2018; Bhat et al., 2021) to cross-dataset generalization. MiDaS (Ranftl et al., 2020; Birkl et al., 2023) enables joint learning from heterogeneous datasets with scale- and shift-invariant objectives, and Depth Anything (Yang et al., 2024a;b) scales this paradigm with pretrained visual encoders and large collections of pseudo-labeled images. Diffusion-based methods offer another route to zero-shot depth estimation (Ke et al., 2025b; Gui et al., 2025; He et al., 2025; Garcia et al., 2025; Pham et al., 2025; Wang et al., 2026a; Song et al., 2026). Using diffusion priors (Rombach et al., 2022; Chen et al., 2024), these methods achieve zero-shot transfer and recover fine-grained details. Beyond affine-invariant relative depth, recent methods have advanced toward zero-shot metric depth estimation, aiming to recover metric-scale geometry across unseen scenes and camera configurations (Hu et al., 2024; Piccinelli et al., 2024; 2025b; Bochkovskiy et al., 2025; Guo et al., 2025; Ganesan et al., 2026). However, depth-space accuracy does not necessarily translate into faithful 3D geometry after back-projection. Recent monocular geometry methods address this issue by explicitly modeling camera geometry or directly predicting 3D point maps (Wang et al., 2025b; Piccinelli et al., 2025a; Wang et al., 2026b; Kong et al., 2026; Xu et al., 2026c).

Video depth estimation introduces temporal context to mitigate the frame-to-frame instability of image predictors. Early methods couple predictions across a video through test-time optimization or global alignment to reduce inconsistency (Luo et al., 2020; Kopf et al., 2021; Zhang et al., 2021), but incur substantial computational cost. Later feed-forward methods avoid this optimization by encoding history through recurrent updates or persistent memory (Zhang et al., 2019; Wang et al.,

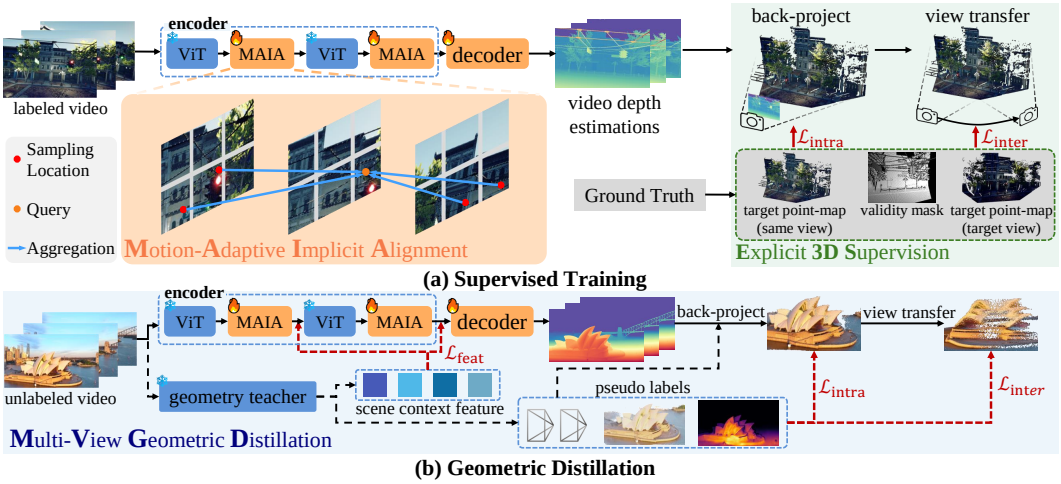


Figure 2: **Method overview.** We introduce G2, a geometry-grounded video depth estimation framework that unifies motion-adaptive cross-frame modeling and explicit geometric supervision. **(a) Supervised Training.** MAIA adaptively aggregates cross-frame evidence, while E3DS back-projects the predicted depth and directly constrains the recovered geometry in 3D space using ground-truth supervision. **(b) Geometric Distillation.** To extend geometry-grounded learning to unlabeled videos, MVGD transfers scene-level geometric context and pseudo labels from a frozen geometry teacher.

2019; Yasarla et al., 2023; Chou et al., 2025). Recent methods transfer strong image-based depth priors to video depth estimation by augmenting pretrained monocular depth models with temporal stabilization modules (Wang et al., 2023; Kuang et al., 2025; Chen et al., 2025; Sobko et al., 2026). Concurrently, video diffusion approaches exploit priors learned by pretrained generators and achieve strong generalization to real-world scenes (Hu et al., 2025; Shao et al., 2025; Yang et al., 2025). Recent works have recognized the importance of geometric structure in video depth estimation. Some leverage feed-forward models (Liu et al., 2026; Xu et al., 2026a) or external correspondences (Wang et al., 2025c; Lai & Vedaldi, 2025; Cho et al., 2025) to establish cross-frame associations, while others back-project depth predictions into 3D and enforce consistency among the resulting geometry (Zhao et al., 2024; Yuan et al., 2025). These advances show the value of cross-frame correspondence modeling and geometric constraints, but either retain geometry-specific computation at inference or use geometry primarily to regularize consistency among predictions. In contrast, G2 preserves depth-only inference without auxiliary reconstruction models or geometric inputs, while directly supervising the recovered 3D geometry against geometric references during training.

3 Methodology

Given a monocular video $\mathcal{I} = \{\mathbf{I}_t\}_{t=1}^{T}$, where $\mathbf{I}_t$ denotes the t-th frame and T is the number of frames, G2 estimates a scale-invariant depth sequence $\mathcal{D}^{\text{pre}} = \{\mathbf{D}_t^{\text{pre}}\}_{t=1}^{T}$, which is defined up to a single global scale relative to the ground truth.

As illustrated in Fig. 2, G2 adopts an encoder-decoder architecture. The early encoder layers process each frame independently to extract dense visual tokens, while the later layers incorporate *Motion-Adaptive Implicit Alignment* (§3.1) to aggregate geometric information from corresponding scene content across frames. The resulting spatiotemporal features are decoded into depth maps, whose back-projected 3D geometry is directly constrained by *Explicit 3D Supervision* (§3.2). During training, *Multi-View Geometric Distillation* (§3.3) further extends geometric supervision to unlabeled videos by distilling scene-level knowledge from a pretrained multi-view reconstruction teacher.

3.1 Motion-Adaptive Implicit Alignment

To associate spatially displaced observations of the same scene point, we introduce MAIA. In contrast to fixed-location temporal attention and methods relying on externally estimated correspondences, MAIA determines its sampling locations directly from the query embeddings.

Given a feature volume consisting of T frames and S spatial tokens per frame, we obtain query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ embeddings through linear projections, each with feature dimension C. For a token indexed by (t, s), let $\mathbf{q}_{t,s} \in \mathbb{R}^C$ denote its query embedding, where $t \in \{1, \dots, T\}$ indexes the frame and $s \in \{1, \dots, S\}$ indexes the spatial token within that frame. Its reference coordinate is $\mathbf{r}_{t,s} = (x_s, y_s, t)$. Under the multi-head attention formulation, this coordinate serves as the sampling origin for all H attention heads. Specifically, each head $h \in \{1, \dots, H\}$ is associated with P preset anchors $\{\overline{\boldsymbol{\delta}}_{h,p}\}_{p=1}^{P}$ and an offset predictor $f_{\text{off}}^h : \mathbb{R}^C \to \mathbb{R}^{P\times 3}$ that predicts query-conditioned residuals. The resulting sampling coordinates are

$$\mathbf{c}_{t,s,h,p} = \mathbf{r}_{t,s} + \overline{\boldsymbol{\delta}}_{h,p} + \left[f_{\text{off}}^h(\mathbf{q}_{t,s})\right]_p, \tag{1}$$

where $[\cdot]_p \in \mathbb{R}^3$ selects the residual for the p-th location.

We retrieve key and value features at continuous sampling locations by trilinear interpolation $(\widetilde{\mathbf{k}}_{t,s,h,p}, \widetilde{\mathbf{v}}_{t,s,h,p}) = \text{TriInterp}(\mathbf{K}^h, \mathbf{V}^h, \mathbf{c}_{t,s,h,p})$. Let $\Omega_{t,s,h} \subseteq \{1, \dots, P\}$ denote valid sampling indices within the feature volume, we compute query-key similarities and normalize them over valid locations to produce attention weights $a_{t,s,h,p}$ that aggregate the sampled values into $\mathbf{o}_{t,s}^h$:

$$\begin{aligned} a_{t,s,h,p} &= \text{Softmax}_{p\in\Omega_{t,s,h}}\left((\mathbf{q}_{t,s}^h)^\top \widetilde{\mathbf{k}}_{t,s,h,p}/\sqrt{d}\right), \\ \mathbf{o}_{t,s}^h &= \sum_{p\in\Omega_{t,s,h}} a_{t,s,h,p}\widetilde{\mathbf{v}}_{t,s,h,p}. \end{aligned} \tag{2}$$

Here, d is the head dimension. The offsets determine where to sample, and attention weights determine how to aggregate the valid sampled features.

3.2 Explicit 3D Supervision

To supervise 3D geometry beyond per-frame depth errors, Explicit 3D Supervision (E3DS) comprises an Intra-Frame Point-Map Loss and an Inter-Frame Projection Loss, which target per-frame scene structure and cross-view geometric consistency, respectively. To account for global scale ambiguity, we align the predicted depth sequence to the target using a single video-level scale shared across all frames. Specifically, given target depths $\mathbf{D}^{\text{tar}}t$ and their validity mask mt, u for pixel u in frame t, we compute the optimal scale factor $s^\star$ as

$$s^\star = \arg\min_s \sum_{t,u} m_{t,u}\left(s\mathbf{D}_{t,u}^{\text{pre}} - \mathbf{D}_{t,u}^{\text{tar}}\right)^2, \qquad \mathbf{D}_t = s^\star\mathbf{D}_t^{\text{pre}}. \tag{3}$$

Intra-Frame Point-Map Loss. For a pixel u with homogeneous image coordinate $\widetilde{\mathbf{u}} = (u_x, u_y, 1)^\top$, we back-project the aligned depth into 3D using the target intrinsic matrix $\mathbf{K}_t^{\text{tar}}$ as $\mathbf{P}_{t,u} = \mathbf{D}_{t,u}(\mathbf{K}_t^{\text{tar}})^{-1}\widetilde{\mathbf{u}}$. We compare the resulting point map with the target $\mathbf{P}_t^{\text{tar}}$ in the coordinate system of camera t using a weighted robust loss:

$$\mathcal{L}_{\text{intra}} = \frac{\sum_{t,u} m_{t,u}\omega_{t,u}\rho_{\text{SL1}}\left(\mathbf{P}_{t,u} - \mathbf{P}_{t,u}^{\text{tar}}\right)}{\sum_{t,u} m_{t,u}\omega_{t,u}}. \tag{4}$$

Here, ρ_{SL1} denotes Smooth L1 penalties and $\omega_{t,u}$ is the point-wise supervision confidence. For labeled data, we set $\omega_{t,u} \propto (\mathbf{D}_{t,u}^{\text{tar}})^{-1}$ to reduce the relative contribution of distant points.

Inter-Frame Projection Loss. The intra-frame point-map loss supervises each frame independently and does not explicitly enforce consistency across views. To explicitly enforce consistency across views, we introduce an inter-frame projection loss that transforms source-frame predictions into a target view and compares them with the target-view geometry. For an ordered frame pair (i, j), the target camera poses define the relative transformation $(\mathbf{R}_{ji}, \mathbf{t}_{ji})$, and each source point is projected onto frame j as $\widehat{\mathbf{v}}_{i\to j,u} = \pi\big(\mathbf{K}_j^{\text{tar}}(\mathbf{R}_{ji}\mathbf{P}_{i,u} + \mathbf{t}_{ji})\big)$, where π denotes perspective division. Bilinear interpolation maps visible projected points to the target grid, yielding $\widetilde{\mathbf{P}}_{i\to j,v}$ in the coordinate system of camera j for comparison with $\mathbf{P}_{j,v}^{\text{tar}}$.

Independent motion and occlusion may invalidate projected correspondences. We therefore construct a geometric validity mask to exclude invalid correspondences by projecting the target point map $\mathbf{P}_i^{\text{tar}}$

to obtain $\widetilde{\mathbf{P}}^{\text{tar}}_{i\to j,v}$. Its consistency with $\mathbf{P}^{\text{tar}}_{j,v}$ is measured by $\text{Rel}_{\text{3D}}(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_2/(\|\mathbf{p}\|_2+\epsilon)$, based on which we define the geometric validity mask as:

$$G_{i\to j,v} = \mathbb{1}\left[\text{Rel}_{\text{3D}}\left(\widetilde{\mathbf{P}}^{\text{tar}}_{i\to j,v}, \mathbf{P}^{\text{tar}}_{j,v}\right) \le \tau_{\text{geo}}\right], \tag{5}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function and τ_{geo} is the geometric consistency threshold.

We sample a set of ordered frame pairs from the T input frames, denoted by $\mathcal{C}$. Over these pairs, the cross-frame projection loss is

$$\mathcal{L}_{\text{inter}} = \frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}} \frac{\sum_v G_{i\to j,v}\omega_{i\to j,v}\rho_{\text{SL1}}\left(\widetilde{\mathbf{P}}_{i\to j,v} - \mathbf{P}^{\text{tar}}_{j,v}\right)}{\sum_v G_{i\to j,v}\omega_{i\to j,v}}. \tag{6}$$

Here, $\omega_{i\to j,v}$ weights projected supervision reliability and equals 1 for valid projections with ground-truth geometry.

Supervised Training Objective. We combine the intra-frame and inter-frame constraints into the E3DS objective $\mathcal{L}_{\text{3D}}$, and further incorporate depth-space supervision to form the overall objective for supervised training.

$$\begin{aligned} \mathcal{L}_{\text{3D}} &= \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}}, \\ \mathcal{L}_{\text{sup}} &= \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{3D}}. \end{aligned} \tag{7}$$

Here, $\mathcal{L}_{\text{sup}}$ jointly optimizes depth and reconstructed 3D geometry during supervised training, where λ_{intra} and λ_{inter} balance intra-frame accuracy and inter-frame consistency. $\mathcal{L}_{\text{depth}}$ follows the supervision proposed by MiDaS (Birkl et al., 2023).

3.3 Multi-View Geometric Distillation

E3DS relies on paired depth and camera annotations, for which labeled data are scarce. To scale geometric supervision to unlabeled videos, we distill geometric knowledge from a frozen multi-view reconstruction teacher.

Feature Distillation. High-level reconstruction features integrate geometric information across views, providing scene-level context for globally coherent 3D reconstruction. In contrast, their location-dependent variations may provide less reliable fine-grained details. Therefore, we distill only the shared scene-level context, transferring geometric structure while preserving the student's ability to recover precise local depth details.

We denote the teacher feature at spatial token s in frame t by $\mathbf{F}^{\text{tea}}_{t,s}$. Rather than imposing direct token-wise feature matching, we extract their scene-level context and transfer it to aligned student features $\mathbf{F}^{\text{stu}}_{t,s}$ constructed from the encoder features after the MAIA blocks. For both representations $b \in \{\text{tea}, \text{stu}\}$, we spatially average the features within each frame to obtain a frame-wise descriptor $\mathbf{g}^b_t = \frac{1}{N}\sum_{s=1}^{N}\mathbf{F}^b_{t,s}$, where N is the number of spatial tokens per frame on the teacher feature grid. We compute pairwise similarities between descriptors and normalize them to obtain the relational distribution $\mathbf{A}^b \in \mathbb{R}^{T\times T}$. The cross-view context loss aligns student and teacher distributions via KL divergence:

$$\mathcal{L}_{\text{feat}} = D_{\text{KL}}\left(A^{\text{tea}} \,\|\, A^{\text{stu}}\right). \tag{8}$$

Distillation Objective. In addition, we use the teacher's reconstruction outputs as output-level supervision. For unlabeled video, the teacher provides target depth $\mathbf{D}^{\text{tar}}_t$, point maps $\mathbf{P}^{\text{tar}}_t$, camera intrinsics $\mathbf{K}^{\text{tar}}_t$, and relative camera transformations $(\mathbf{R}_{ji}, \mathbf{t}_{ji})$, with each point map expressed in its corresponding camera coordinate system. The associated teacher confidences are used as supervision weights $\omega_{t,u}$ and $\omega_{i\to j,v}$. Together, these estimated quantities serve as pseudo ground truth in the formulation of $\mathcal{L}_{\text{3D}}$ in §3.2, yielding the teacher-guided geometric loss $\mathcal{L}^{\text{tea}}_{\text{3D}}$. The objective for distillation is

$$\mathcal{L}_{\text{distill}} = \lambda^{\text{tea}}_{\text{3D}}\mathcal{L}^{\text{tea}}_{\text{3D}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}, \tag{9}$$

where $\lambda^{\text{tea}}_{\text{3D}}$ and λ_{feat} balance label and feature distillation.

Table 1: **Quantitative results for zero-shot video depth estimation.** Best and second-best results are highlighted in **bold** and *italics*. (see §4.2)

Alignment	Method	Avg. Rank ↓	Bonn		ScanNet		Sintel	
			AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
SSI	RollingDepth	6.7	0.056	0.980	0.084	0.934	0.370	0.591
	DepthCrafter	6.8	0.060	0.975	0.100	0.912	0.359	0.677
	DepthAnyVideo	5.5	0.056	0.977	0.082	0.945	0.275	0.674
	VDA	5.8	0.053	0.972	0.076	0.954	0.301	0.638
	GeometryCrafter	3.3	0.048	0.980	0.069	0.965	0.289	0.713
	GemDepth	*2.8*	0.051	0.982	0.069	0.959	**0.160**	**0.811**
	PPVD	3.0	**0.032**	**0.989**	**0.029**	**0.990**	0.528	0.662
	G2	**2.0**	*0.048*	*0.984*	*0.068*	*0.965*	*0.214*	*0.750*
SI	RollingDepth	7.7	0.800	0.357	0.603	0.446	0.519	0.394
	DepthCrafter	5.2	0.467	0.408	0.281	0.658	1.549	0.450
	DepthAnyVideo	6.7	0.777	0.360	0.378	0.606	1.283	0.426
	VDA	4.7	0.400	0.492	0.361	0.563	0.425	0.441
	GeometryCrafter	*1.7*	**0.048**	*0.981*	*0.076*	**0.961**	*0.262*	*0.594*
	GemDepth	5.3	0.569	0.406	0.373	0.560	0.431	0.572
	PPVD	3.5	0.192	0.750	0.216	0.761	0.487	0.516
	G2	**1.3**	*0.051*	**0.984**	**0.076**	*0.949*	**0.198**	**0.700**

3.4 Implementation Details

We initialize the encoder and decoder parameters with pretrained weights from Video Depth Anything (Chen et al., 2025). The pretrained encoder remains frozen, while three trainable MAIA blocks are inserted after encoder blocks 15, 19, and 23, each with $P = 4$ sampling locations per head. We use Pi3 (Wang et al., 2026c) as the frozen multi-view reconstruction teacher and extract teacher features from the penultimate layer of its multi-view aggregator, where multi-view information has been extensively integrated while remaining independent of the task-specific prediction heads. For feature distillation, two MLPs generate gates to fuse selected student features, and another MLP projects the fusion to the teacher feature dimension.

Model training consists of two stages: *supervised learning* and *geometric distillation*. In the supervised training stage, we train the model on labeled data from TartanAir (Wang et al., 2020), Virtual KITTI (Cabon et al., 2020), MVS-Synth (Huang et al., 2018), PointOdyssey (Zheng et al., 2023), and UnrealStereo4K (Tosi et al., 2021) using the supervised objective $\mathcal{L}_{\text{sup}}$ from Eq. 7. In the subsequent geometric distillation stage, we retain a small set of synthetic-data batches under the same supervised objective $\mathcal{L}_{\text{sup}}$. These are alternated with unlabeled batches from OpenVid-1M (Nan et al., 2025) and Kinetics-700-2020 (Smaira et al., 2020), which are optimized using the distillation objective $\mathcal{L}_{\text{distill}}$ in Eq. 9.

Each training stage runs for 20 epochs on four NVIDIA RTX 3090 GPUs using AdamW with a weight decay of 0.01 and an effective batch size of 16. Within each stage, the learning rate is increased to 5×10^{-5} during the first epoch, held constant throughout the second epoch, and decayed according to a cosine schedule thereafter. We set $\tau_{\text{geo}} = 0.02$, $\lambda_{\text{intra}} = 10$, $\lambda_{\text{inter}} = 0.5$, and $\lambda_{\text{feat}} = 0.25$ by default across the experiments.

4 Experiments

4.1 Evaluation Protocol

Datasets. For video depth estimation, we follow prior protocols (Hu et al., 2025; Chen et al., 2025) and evaluate on Sintel (Butler et al., 2012), ScanNet (Dai et al., 2017), and Bonn (Palazzolo et al., 2019) datasets. Following the reconstruction evaluation settings of prior works (Wang et al., 2025a; 2026c), we evaluate geometric reconstruction on 7-Scenes (Shotton et al., 2013) and NRGBD (Azinović et al., 2022). None of these datasets is used for training.

Table 3: **Quantitative results for geometric reconstruction.** Median metrics are reported under dense-view evaluation. Best values are shown in **bold**. (see §4.3)

Method	7-Scenes						NRGBD					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	SSI	SI	SSI	SI	SSI	SI	SSI	SI	SSI	SI	SSI	SI
RollingDepth	0.035	0.318	0.014	0.034	0.765	0.596	0.097	0.515	0.025	0.056	0.904	0.695
DepthCrafter	0.086	0.313	0.028	0.032	0.749	0.618	0.209	0.466	0.044	0.050	0.859	0.748
DepthAnyVideo	0.047	0.113	0.017	0.021	0.768	0.673	0.096	0.233	0.023	0.041	0.936	0.830
VDA	0.033	0.122	0.014	0.049	0.795	0.739	0.094	0.285	0.026	0.060	0.961	0.864
GeometryCrafter	0.036	0.030	0.016	0.013	0.796	0.778	0.082	0.095	0.022	0.023	0.957	0.950
GemDepth	0.031	0.087	0.014	0.019	0.798	0.698	0.092	0.235	0.026	0.041	0.960	0.865
PPVD	**0.020**	0.321	**0.009**	0.049	0.735	0.464	0.079	0.616	0.019	0.113	0.850	0.491
G2	0.024	**0.021**	0.013	**0.011**	**0.802**	**0.794**	**0.057**	**0.057**	**0.019**	**0.020**	**0.962**	**0.962**

Metrics. Following prior studies (Chen et al., 2025; Wang et al., 2026c), we evaluate depth estimation using Absolute Relative Error (AbsRel), δ_1 accuracy, and Temporal Alignment Error (TAE), and assess 3D reconstruction using Accuracy (Acc.), Completion (Comp.), and Normal Consistency (N.C.). All metrics are reported under scale-and-shift-invariant (SSI) and scale-invariant (SI) alignment.

4.2 Zero-shot Video Depth Estimation

To evaluate G2, we compare it with recent video depth estimation methods, including RollingDepth (Ke et al., 2025a), DepthCrafter (Hu et al., 2025), DepthAnyVideo (Yang et al., 2025), Video Depth Anything (VDA) (Chen et al., 2025), GeometryCrafter (Xu et al., 2025), GemDepth (Liu et al., 2026), and Pixel-Perfect Video Depth (PPVD) (Xu et al., 2026b). Table 1 reports zero-shot accuracy, with Avg. Rank computed over AbsRel and δ_1 across all benchmarks. Under SSI alignment, G2 achieves the best average rank of $\mathbf{2.0}$, demonstrating consistently competitive depth estimation across datasets. This result is consistent with our geometry-grounded objective, which encourages the predicted depth sequence to preserve the underlying 3D structure up to a global scale. The additional shift degree of freedom in SSI may offer limited benefit to such predictions, while providing greater compensation for methods with larger affine ambiguity, thereby narrowing G2's relative advantage on individual SSI metrics. Under the stricter SI alignment, its advantage becomes more pronounced, achieving the best average rank of $\mathbf{1.3}$ and ranking first or second on all six reported metrics. Notably, G2 achieves a δ_1 of $\mathbf{0.700}$ on Sintel under SI alignment, outperforming the best competing method by $\mathbf{10.5}$ percentage points.

We further evaluate temporal consistency beyond per-frame depth accuracy. Table 2 reports TAE under both alignments across all three benchmarks, measuring adjacent-frame consistency. G2 achieves the strongest overall temporal consistency, ranking first in three of the six settings and attaining the best average rank of $\mathbf{2.3}$ across all reported TAE values. Under SI alignment, G2 substantially reduces TAE over VDA by $\mathbf{53.7}\%$ on Bonn and $\mathbf{35.9}\%$ on ScanNet, while further reducing TAE by $\mathbf{7.6}\%$ on Sintel under SSI alignment.

Table 2: **Quantitative results for temporal consistency.** TAE (↓) is reported.

Method	Bonn		ScanNet		Sintel	
	SSI	SI	SSI	SI	SSI	SI
RollingDepth	2.339	6.723	1.180	3.118	59.207	35.043
DepthCrafter	**2.229**	5.492	0.652	*1.139*	59.909	39.002
VDA	2.383	5.483	*0.549*	1.261	*55.214*	30.872
GemDepth	*2.247*	5.985	**0.537**	1.241	57.948	48.556
PPVD	2.489	*4.307*	0.790	1.424	58.630	**19.763**
G2	2.378	**2.539**	0.795	**0.808**	**51.024**	*28.302*

4.3 3D Reconstruction

To assess 3D structure beyond depth maps, we back-project aligned depth predictions using ground-truth intrinsics and transform them into a common coordinate system with ground-truth extrinsics.

As shown in Table 3, G2 demonstrates strong geometric reconstruction performance compared with previous generative and discriminative methods. On NRGBD, G2 achieves an accuracy error of $\mathbf{0.057}$ under both alignments, reducing the previous best results by $\mathbf{28.3}\%$ and $\mathbf{39.6}\%$. It further

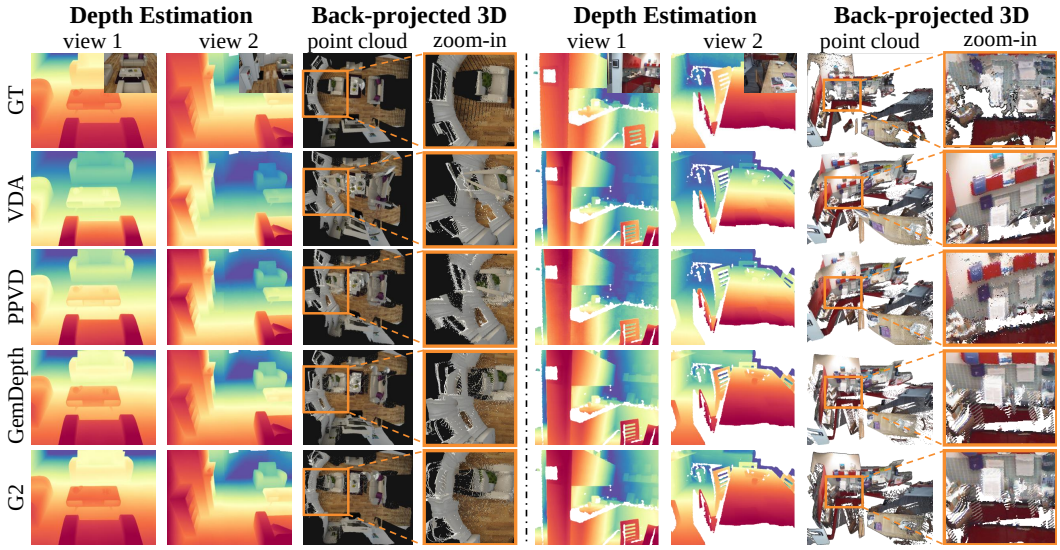


Figure 3: **Qualitative comparison** of depth estimation and geometric reconstruction. (see §4.3)

Table 4: **Comparison of cross-frame modeling mechanisms.**

Method	7-Scenes SSI Acc. ↓	7-Scenes SSI Comp. ↓	7-Scenes SI Acc. ↓	7-Scenes SI Comp. ↓	Bonn SSI δ_1 ↑	Bonn SI δ_1 ↑
None	0.031	0.016	0.029	0.015	0.974	0.976
TA	0.027	0.013	0.024	0.012	0.981	0.979
MAIA	**0.025**	**0.013**	**0.022**	**0.011**	**0.981**	**0.980**

Table 5: **Ablation study on learnable residual offsets in MAIA.** The offsets are predicted by f^h_{off} in Eq. 1.

Offsets	7-Scenes SSI Acc. ↓	7-Scenes SSI Comp. ↓	7-Scenes SI Acc. ↓	7-Scenes SI Comp. ↓	Bonn SSI δ_1 ↑	Bonn SI δ_1 ↑
✗	0.028	0.017	0.024	0.013	0.973	0.967
✓	**0.025**	**0.013**	**0.022**	**0.011**	**0.981**	**0.980**

achieves the best normal consistency under both alignments and the best SI-aligned completion on NRGBD, while ranking first or second across all reconstruction metrics on 7-Scenes. Notably, these geometric gains persist under the stricter SI alignment, indicating that G2 recovers more accurate scene structure without relying on an additional shift correction.

Qualitative Results. Figure 3 further illustrates the importance of grounding video depth predictions in the shared 3D scene underlying different views. Although competing methods produce visually plausible depth maps, back-projecting their predictions from multiple views into a common 3D space reveals structural distortions and cross-view misalignments, which are particularly evident in the highlighted regions. In contrast, G2 preserves detailed depth predictions while yielding better-aligned surfaces and more coherent scene geometry across views. These qualitative observations are consistent with the quantitative results, showing that G2 preserves competitive depth estimation performance while substantially improving geometric reconstruction.

4.4 Ablation Study

Motion-Adaptive Implicit Alignment. We report ablations on 7-Scenes for reconstruction and Bonn for depth estimation. To isolate the benefit of MAIN over conventional temporal aggregation, we replace it with temporal attention while keeping the remaining architecture unchanged. Table 4 compares their depth accuracy and reconstruction quality under the same evaluation setting, where TA denotes temporal attention. Compared with temporal attention, MAIA reduces the 7-Scenes accuracy errors by **7.5**% and **6.6**%. These results show that MAIA improves the geometry recovered from predicted depth, beyond maintaining per-frame depth accuracy.

Table 5 further isolates the contribution of the learnable residual offset in MAIA. Compared with using only the preset anchor, learnable residual offset reduce the SSI/SI accuracy errors by **12.2**%/**8.3**% and the corresponding completion errors by **24.2**%/**12.2**%. It also improve Bonn δ_1 by **0.8**/**1.3** percentage points, respectively. These gains support adapting the cross-frame sampling locations through query-conditioned offset to improve both depth prediction and recovered geometry.

Table 6: **Ablations on the number of sampled locations P in MAIA.**

P	7-Scenes SSI Acc. ↓	7-Scenes SSI Comp. ↓	7-Scenes SI Acc. ↓	7-Scenes SI Comp. ↓	Bonn SSI δ_1 ↑	Bonn SI δ_1 ↑
2	0.028	0.015	0.024	0.013	**0.982**	**0.981**
4	**0.025**	**0.013**	**0.022**	**0.011**	0.981	0.980
8	0.026	0.013	0.024	0.012	0.981	0.980

Table 7: **Ablations on components of Explicit 3D Supervision.**

$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{inter}}$	7-Scenes SSI Acc. ↓	7-Scenes SSI Comp. ↓	7-Scenes SI Acc. ↓	7-Scenes SI Comp. ↓	Bonn SSI δ_1 ↑	Bonn SI δ_1 ↑
✗	✗	0.031	0.014	0.029	0.013	0.981	**0.982**
✓	✗	0.027	0.013	0.024	0.012	0.981	0.981
✗	✓	0.026	0.014	0.024	0.012	**0.982**	0.981
✓	✓	**0.025**	**0.013**	**0.022**	**0.011**	0.981	0.980

Table 8: **Ablations on the inter-frame projection loss weight λ_{inter}.** The default setting is $\lambda_{\text{inter}} = 0.5$.

λ_{inter}	7-Scenes SSI Acc. ↓	7-Scenes SSI Comp. ↓	7-Scenes SI Acc. ↓	7-Scenes SI Comp. ↓	Bonn SSI δ_1 ↑	Bonn SI δ_1 ↑
0.25	0.027	0.013	0.024	0.012	**0.982**	**0.981**
0.5	**0.025**	**0.013**	**0.022**	**0.011**	0.981	0.980
1.0	0.026	0.013	0.023	0.012	0.979	0.979

Table 9: **Ablations on the supervision components of Multi-View Geometric Distillation.**

$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{3D}}^{\text{tea}}$	7-Scenes SSI Acc. ↓	7-Scenes SSI Comp. ↓	7-Scenes SI Acc. ↓	7-Scenes SI Comp. ↓	Bonn SSI δ_1 ↑	Bonn SI δ_1 ↑
✗	✗	0.025	0.013	0.022	0.011	0.981	0.980
✗	✓	0.025	0.013	0.022	0.011	**0.984**	0.984
✓	✓	**0.024**	**0.013**	**0.021**	**0.011**	0.984	**0.984**

Table 6 examines the number of sampling locations P in MAIA. Increasing P from 2 to 4 reduces the reconstruction errors by $\mathbf{9.3}\%$–$\mathbf{15.0}\%$, while increasing P further to 8 provides no additional improvement. Across the tested settings, Bonn δ_1 varies by at most $\mathbf{0.13}$ percentage points, remaining essentially unchanged. These results indicate that a small set of adaptive sampling locations is sufficient for effective geometric reconstruction, while depth accuracy remains insensitive.

Explicit 3D Supervision. We ablate the intra-frame point-map loss and inter-frame projection loss to assess their individual contributions, with the results reported in Table 7. Both losses independently improve geometric reconstruction, while their combination yields the most consistent gains, reducing the median accuracy error by $\mathbf{0.006}$ and $\mathbf{0.007}$ under SSI and SI alignment compared with using neither loss. Meanwhile, the depth accuracy on Bonn remains nearly unchanged, showing that E3DS improves the recovered 3D geometry while maintaining depth estimation performance.

Table 8 examines sensitivity to the inter-frame projection weight λ_{inter}. Across the tested range, all three settings achieve lower errors on all four reconstruction metrics than the depth-only baseline in Table 7, indicating that the geometric gains from inter-frame supervision remain effective over a broad range of weights. Increasing λ_{inter} from 0.25 to 0.5 improves geometric reconstruction and reduces reconstruction metrics under both alignments, including an $\mathbf{8.2}\%$ reduction in Acc. under SSI alignment. However, a larger weight of 1.0 provides no additional improvement and slightly degrades depth accuracy. We therefore set $\lambda_{\text{inter}} = 0.5$, which achieves the best trade-off between geometric reconstruction and depth accuracy among the tested settings.

Multi-View Geometric Distillation. Table 9 evaluates the contributions of feature distillation and teacher-guided geometric supervision on real videos. Combining both components improves all six reported metrics over synthetic supervision only: the four reconstruction errors decrease by $\mathbf{0.5}\%$–$\mathbf{2.4}\%$, and Bonn δ_1 increases by $\mathbf{0.28/0.44}$ percentage points for SSI/SI, respectively.

5 Conclusion

This paper present G2, a novel framework for achieving both temporally and geometrically consistent video depth estimation. G2 addresses three key limitations of existing methods. First, motion-adaptive implicit alignment adapts cross-frame interactions to camera and object motion by dynamically locating and aggregating sparse query-relevant features. Second, explicit 3D supervision directly constrains the recovered geometry through intra-frame point-map and inter-frame projection losses. Third, multi-view geometric distillation extends geometric supervision to unlabeled videos by transferring geometric knowledge from a frozen reconstruction teacher. Extensive experiments on 10 benchmarks show that G2 outperforms existing state-of-the-art approaches on both zero-shot video depth estimation and geometric reconstruction.

AI USE STATEMENT

Generative AI tools were used only for language editing to improve the clarity and readability of the manuscript, and for limited assistance with code formatting and debugging. All core research ideas, methodological designs, experimental decisions, analyses, and scientific conclusions were independently developed by the authors. All AI-assisted text was manually reviewed, and all AI-assisted code was inspected and, where necessary, modified by the authors. The use of generative AI tools was strictly limited to the above purposes, in accordance with ICLR 2027 AI Policy. The authors take full responsibility for the final content of this work.

REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *CVPR*, pp. 6290–6301, 2022.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pp. 4009–4018, 2021.

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.

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 *ICLR*, pp. 75602–75637, 2025.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pp. 611–625, 2012.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, pp. 57611–57640, 2024.

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*, pp. 22831–22840, 2025.

Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, pp. 7063–7072, 2019.

Seokju Cho, Jiahui Huang, Seungryong Kim, and Joon-Young Lee. Seurat: From moving points to depth. In *CVPR*, pp. 7211–7221, 2025.

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. In *ICCV*, pp. 9638–9648, 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*, pp. 5828–5839, 2017.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pp. 2650–2658, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pp. 2002–2011, 2018.

Girish Chandar Ganesan, Yuliang Guo, Liu Ren, and Xiaoming Liu. Unidac: Universal metric depth estimation for any camera. In *CVPR*, pp. 26953–26963, 2026.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *WACV*, pp. 753–762, 2025.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *AAAI*, pp. 3203–3211, 2025.

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 *CVPR*, pp. 26996–27006, 2025.

Leezy Han, Seunggyu Kim, Dongseok Shim, and Hyeonbeom Lee. Ptc-depth: Pose-refined monocular depth estimation with temporal consistency. *arXiv preprint arXiv:2604.01791*, 2026.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *ICLR*, pp. 89454–89467, 2025.

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 Trans. Pattern Anal. Mach. Intell.*, 46(12):10579–10596, 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*, pp. 2005–2015, 2025.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, pp. 2821–2830, 2018.

Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. In *CVPR*, pp. 7233–7243, 2025a.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025b.

Numair Khan, Eric Penner, Douglas Lanman, and Lei Xiao. Temporally consistent online depth estimation using point-based fusion. In *CVPR*, pp. 9119–9129, 2023.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. Moge-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement. *arXiv preprint arXiv:2607.17967*, 2026.

Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, pp. 1611–1621, 2021.

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*, pp. 17660–17670, 2025.

Zihang Lai and Andrea Vedaldi. Tracktention: Leveraging point tracking to attend videos faster and better. In *CVPR*, pp. 22809–22819, 2025.

Yuecheng Liu, Junda Cheng, Longliang Liu, Wenjing Liao, Hanrui Cheng, Yuzhou Wang, and Xin Yang. Gemdepth: Geometry-embedded features for 3d-consistent video depth. *arXiv preprint arXiv:2605.10525*, 2026.

Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Trans. Graph.*, 39(4):71–1, 2020.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *ICLR*, pp. 1045–1064, 2025.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *IROS*, pp. 7855–7862, 2019.

Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *CVPR*, pp. 17060–17069, 2025.

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*, pp. 10106–10116, 2024.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *CVPR*, pp. 1028–1039, 2025a.

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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025b.

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 Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

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*, pp. 22841–22852, 2025.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, pp. 2930–2937, 2013.

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

Ivan Sobko, Hayko Riemenschneider, Markus Gross, and Christopher Schroers. Stabledpt: Temporal stable monocular video depth estimation. *arXiv preprint arXiv:2601.02793*, 2026.

Ziyang Song, Zerong Wang, Bo Li, Hao Zhang, Ruijie Zhu, Li Liu, Peng-Tao Jiang, and Tianzhu Zhang. Depthmaster: Taming diffusion models for monocular depth estimation. *IEEE Trans. Circuits Syst. Video Technol.*, 2026.

Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *CVPR*, pp. 8938–8948, 2021.

Jiyuan Wang, Chunyu Lin, Lang Nie, Jing He, Haodong Li, Kang Liao, Yao Zhao, et al. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *NeurIPS*, 38:79901–79930, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, pp. 10510–10522, 2025a.

Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *CVPR*, pp. 5555–5564, 2019.

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*, pp. 5261–5271, 2025b.

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. *NeurIPS*, 38:35928–35959, 2026b.

Shizun Wang, Zhenxiang Jiang, Xingyi Yang, and Xinchao Wang. C4d: 4d made from 3d through dual correspondences. In *ICCV*, pp. 7570–7580, 2025c.

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*, pp. 4909–4916, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, pp. 10481–10497, 2026c.

Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *ACM MM*, pp. 6347–6358, 2022.

Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *ICCV*, pp. 9466–9476, 2023.

Gangwei Xu, Haotong Lin, Hongcheng Luo, Haiyang Sun, Bing Wang, Guang Chen, Sida Peng, Hangjun Ye, and Xin Yang. Pixel-perfect visual geometry estimation. *arXiv preprint arXiv:2601.05246*, 2026a.

Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi_, Haiyang Sun, Bing Wang, et al. Pixel-perfect depth with semantics-prompted diffusion transformers. *NeurIPS*, 38:174731–174755, 2026b.

Haofei Xu, Rundi Wu, Philipp Henzler, Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Marc Pollefeys, Andreas Geiger, Federico Tombari, and Michael Niemeyer. Pointdit: Pixel-space diffusion for monocular geometry estimation. *arXiv preprint arXiv:2607.02515*, 2026c.

Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *ICCV*, pp. 6632–6644, 2025.

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*, pp. 97335–97349, 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*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 37:21875–21911, 2024b.

Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. Mamo: Leveraging memory and attention for monocular video depth estimation. In *ICCV*, pp. 8754–8764, 2023.

Chengbo Yuan, Geng Chen, Li Yi, and Yang Gao. Self-supervised monocular 4d scene reconstruction for egocentric videos. In *ICCV*, pp. 8863–8874, 2025.

Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *ICCV*, pp. 1725–1734, 2019.

Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM Trans. Graph.*, 40(4):1–12, 2021.

Haimei Zhao, Jing Zhang, Zhuo Chen, Bo Yuan, and Dacheng Tao. On robust cross-view consistency in self-supervised monocular depth estimation. *Machine Intelligence Research*, 21(3):495–513, 2024.

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*, pp. 19798–19808, 2023.

A Evaluation Metrics

A.1 Depth Evaluation Metrics

Let $\widetilde{D}_i$ and D_i^{gt} denote the aligned prediction and ground-truth depth at a valid pixel i, and let M be the number of valid pixels. Absolute Relative Error is defined as

$$\mathrm{AbsRel} = \frac{1}{M}\sum_{i=1}^{M}\frac{\left|\widetilde{D}_i - D_i^{\mathrm{gt}}\right|}{D_i^{\mathrm{gt}}}. \tag{10}$$

The δ_1 accuracy is

$$\delta_1 = \frac{1}{M}\sum_{i=1}^{M}\mathbb{1}\left[\max\left(\frac{\widetilde{D}_i}{D_i^{\mathrm{gt}}}, \frac{D_i^{\mathrm{gt}}}{\widetilde{D}_i}\right) < 1.25\right]. \tag{11}$$

We follow Depth Any Video (Yang et al., 2025) for Temporal Alignment Error. For a sequence of N aligned depth maps, let $f(\widetilde{\mathbf{D}}^k, \mathbf{T}_{k+1\leftarrow k})$ denote back-projecting the depth in frame k, transforming the resulting 3D points to frame $k+1$ using the relative camera transformation $\mathbf{T}_{k+1\leftarrow k}$, and reprojecting them onto frame $k{+}1$. The reverse direction uses $\mathbf{T}_{k\leftarrow k+1} = \mathbf{T}_{k+1\leftarrow k}^{-1}$. The bidirectional temporal error is

$$\begin{aligned}\mathrm{TAE}_{\mathrm{raw}} = \frac{1}{2(N-1)}\sum_{k=1}^{N-1}\Big[&\mathrm{AbsRel}\left(f(\widetilde{\mathbf{D}}^k, \mathbf{T}_{k+1\leftarrow k}), \widetilde{\mathbf{D}}^{k+1}\right)\\ &+\mathrm{AbsRel}\left(f(\widetilde{\mathbf{D}}^{k+1}, \mathbf{T}_{k\leftarrow k+1}), \widetilde{\mathbf{D}}^{k}\right)\Big].\end{aligned} \tag{12}$$

Only valid reprojected pixels are included in each directional error.

A.2 Geometric Reconstruction Metrics

We evaluate geometric reconstruction using Accuracy (Acc), Completeness (Comp), and Normal Consistency (NC). Let $\mathcal{P}$ and $\mathcal{G}$ denote the predicted and ground-truth point clouds, respectively.

Accuracy measures the average distance from each predicted point to its nearest ground-truth point:

$$\mathrm{Acc}(\mathcal{P}, \mathcal{G}) = \frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\min_{\mathbf{g}\in\mathcal{G}}\|\mathbf{p}-\mathbf{g}\|_2. \tag{13}$$

Completeness measures the average distance from each ground-truth point to its nearest predicted point:

$$\mathrm{Comp}(\mathcal{P}, \mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}}\min_{\mathbf{p}\in\mathcal{P}}\|\mathbf{g}-\mathbf{p}\|_2. \tag{14}$$

For Normal Consistency, we define the nearest-neighbor correspondences as

$$\pi_{\mathcal{G}}(\mathbf{p}) = \arg\min_{\mathbf{g}\in\mathcal{G}}\|\mathbf{p}-\mathbf{g}\|_2, \qquad \pi_{\mathcal{P}}(\mathbf{g}) = \arg\min_{\mathbf{p}\in\mathcal{P}}\|\mathbf{g}-\mathbf{p}\|_2. \tag{15}$$

Let $\mathbf{n_p}$ and $\mathbf{n_g}$ denote the unit surface normals at $\mathbf{p}$ and $\mathbf{g}$, respectively. Normal Consistency is computed symmetrically as

$$\mathrm{NC}(\mathcal{P}, \mathcal{G}) = \frac{1}{2}\left(\frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\left|\mathbf{n}_{\mathbf{p}}^{\top}\mathbf{n}_{\pi_{\mathcal{G}}(\mathbf{p})}\right| + \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}}\left|\mathbf{n}_{\mathbf{g}}^{\top}\mathbf{n}_{\pi_{\mathcal{P}}(\mathbf{g})}\right|\right). \tag{16}$$

Lower Acc and Comp indicate more accurate and complete reconstructed geometry, while higher NC indicates better agreement of local surface orientations.

B DISCUSSION

B.1 LIMITATION AND FUTURE WORK

Our work focuses on monocular video depth estimation rather than a complete 3D reconstruction system. Although G2 improves the geometric consistency of predicted depth across frames, directly recovering metrically accurate 3D scenes additionally requires reliable camera pose estimation and absolute scale recovery. Future work could integrate these components into a unified framework for end-to-end video-based 3D reconstruction.

B.2 TERMS OF USE, PRIVACY, AND LICENSE

G2 is available for non-commercial research purposes. The datasets, pretrained models, and third-party software used in this work are subject to their respective licenses and terms of use. Our experiments do not involve the collection of private or personally identifiable information.