INTERP3R: CONTINUOUS-TIME 3D GEOMETRY ESTIMATION WITH FRAMES AND EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, 3D visual foundation models, pioneered by pointmap-based approaches such as DUSt3R, have attracted a lot of interest, achieving impressive accuracy and strong generalization across diverse scenes. However, these methods are inherently limited to recovering scene geometry only at the discrete time instants when images are captured, leaving the scene evolution during the blind time between consecutive frames largely unexplored. We introduce Interp3R, to the best of our knowledge, the first method that enhances pointmap-based models to estimate depth and camera poses at arbitrary time instants. It leverages asynchronous event data to interpolate pointmaps produced by frame-based models, enabling temporally continuous geometric representations. Depth and camera poses are then jointly recovered by aligning the interpolated pointmaps together with those predicted by the underlying frame-based models into a consistent spatial framework. We train Interp3R exclusively on a synthetic dataset, yet demonstrate strong generalization across six datasets, both synthetic and real. Compared with the best two-stage baseline, Interp3R reduces absolute relative depth error by 15%–32% on DSEC and absolute trajectory error by up to 51% on EDS.

1 INTRODUCTION

Recent years have witnessed remarkable progress in 3D visual foundation models. Pointmap-based approaches such as DUSt3R (Wang et al., 2024) and its successors (Zhang et al., 2025; Lu et al., 2025) have demonstrated impressive performance in recovering scene geometry and camera poses from images, while exhibiting strong generalization across diverse environments. These models predict dense 3D pointmaps from input frames and enable various downstream tasks, including the estimation of depth and camera poses. Despite their success, existing methods fundamentally operate on discrete-time frames and can only recover scene geometry at the times of frame capture.

However, real-world scenes evolve continuously over time. The temporal gaps between consecutive frames inevitably introduce blind intervals during which the scene dynamics remain unobserved. This limitation becomes particularly severe in scenarios with fast motion or large frame intervals, where significant geometric changes may occur between frames. As a result, current frame-based methods fail to provide temporally continuous representations of scene geometry.

Event cameras provide a promising solution to this problem. Unlike conventional cameras that capture images at fixed frame rates, event cameras asynchronously record per-pixel brightness changes with microsecond-level temporal resolution (Gallego et al., 2022; Brandli et al., 2014). This allows them to capture fine-grained motion information between frames (Hidalgo-Carrió et al., 2022; Guo et al., 2025), making them well suited for recovering scene evolution in continuous time (Guo & Gallego, 2024; 2025). However, effectively integrating asynchronous event streams with pointmap-based models remains challenging. Events are sparse, noisy, and fundamentally different from frame-based representations, making their direct use in existing 3D visual models non-trivial.

In this paper, we introduce *Interp3R*, a novel method that extends pointmap-based models to continuous-time 3D geometry estimation (see Figure 1). Given two frames processed by a frame-based model, e.g., DUSt3R (Wang et al., 2024), MonST3R (Zhang et al., 2025) or Align3R (Lu et al., 2025), Interp3R leverages the co-captured event stream to interpolate pointmaps at *arbitrary* time instants between the frames. The interpolated pointmaps, together with those predicted from

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.

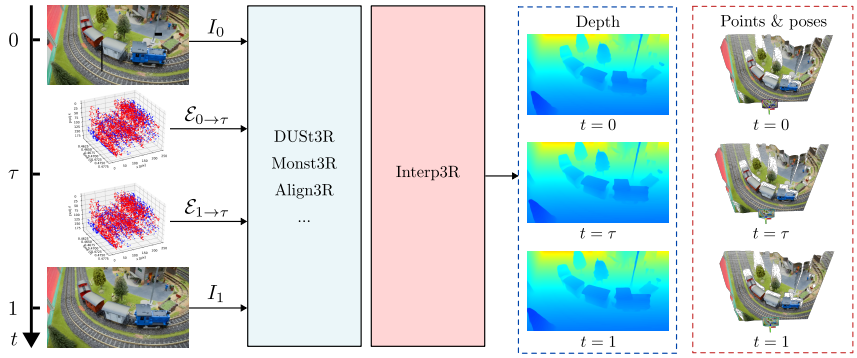

Figure 1: The proposed method takes two images I_0, I_1 and the events between them $\mathcal{E}$ as input (Left), and recovers scene geometry as pointmaps at an arbitrary time instant τ between the images (Right). All available pointmaps can then be used to estimate depth maps and camera poses.

the input frames, are then aligned within a unified spatial framework in a coarse-to-fine manner in order to jointly recover depth and camera poses. This produces robust, progressively refined geometry estimation in continuous time across large temporal gaps.

We train Interp3R exclusively on synthetic data, yet the learned model generalizes well to a wide range of synthetic and real-world datasets. Extensive experiments demonstrate that Interp3R consistently outperforms baselines that follow a two-stage pipeline of 2D video frame interpolation followed by frame-based 3D geometry estimation. Our results show that incorporating event data enables accurate and temporally continuous reconstruction of camera motion and scene geometry.

Our contributions can be summarized as follows:

1. We propose Interp3R, the *first method* that extends pointmap-based models to estimate camera poses and scene geometry in continuous time, enabling the recovery of depth and camera poses at arbitrary time instants. It leverages asynchronous event data to interpolate pointmaps produced by frame-based models, and jointly recovers geometry by aligning the interpolated and frame-based pointmaps in a coarse-to-fine manner.
2. We show that Interp3R is compatible with existing pointmap-based methods (e.g., MonST3R and Align3R), enhancing them to continuous-time space.
3. We train Interp3R exclusively on a synthetic dataset and demonstrate strong generalization across six synthetic and real-world datasets. Extensive experiments show that our method outperforms baseline methods, which follow a two-stage pipeline design of 2D video frame interpolation followed by frame-based 3D geometry estimation.
4. We make the source code publicly available (upon acceptance).

2 Related Work

2.1 3D Geometry Estimation from Images and Videos

Recent years have witnessed significant progress in estimating scene geometry and camera motion from images. Here we focus on reviewing the most relevant pointmap-based methods. A more detailed tabular comparison is provided in Table 4 in the Appendix. DUSt3R (Wang et al., 2024) introduces a pointmap representation and a model that predicts pointmaps from image pairs. It recovers scene geometry and camera poses through a global alignment of all predicted pointmaps, and demonstrates strong generalization across diverse scenes. Building on this idea, MonST3R (Zhang et al., 2025) extends the framework to dynamic scenes by finetuning on dynamic datasets while introducing motion-aware global alignment. Align3R (Lu et al., 2025) further integrates monocular depth priors into the pointmap framework, combining feed-forward depth models and pointmap-based geometry estimation to improve temporal consistency.

Beyond per-frame geometry estimation, recent work has also explored modeling 3D motion between frames. DynaDUSt3R (Jin et al., 2025) augments DUSt3R with motion heads to predict 3D scene flow between two frames at an intermediate query time. Given two input images, the model predicts 3D displacement vectors that describe how points move in space, enabling the estimation of point trajectories across time. However, while DynaDUSt3R can hallucinate sparse 3D motion between frames, it cannot recover camera poses or dense depth maps at intermediate timestamps

In short, existing 3D geometry estimation methods operate at discrete frame timestamps and lack continuous-time depth and pose estimation models to deal with the blind interval between frames.

2.2 Event-based Video Frame Interpolation

Event-based video frame interpolation (EVFI) exploits the high temporal resolution of event cameras to synthesize intermediate frames between sparsely sampled images. TimeLens (Tulyakov et al., 2021) and TimeLens++ (Tulyakov et al., 2022) pioneer this direction by introducing a framework that estimates inter-frame motion from events and synthesizes intermediate frames accordingly. REFID (Sun et al., 2023) performs frame interpolation from blurry video inputs. CBMNet (Kim et al., 2023) improves motion estimation by jointly leveraging image and event modalities through cross-modal asymmetric bidirectional motion fields. TimeLens-XL (Ma et al., 2024) enables arbitrary-time frame interpolation by recursively decomposing large motions into smaller temporal steps. VDM-EVFI (Chen et al., 2025a) explores a generative approach by adapting pretrained video diffusion models for event-based frame interpolation, benefiting from large-scale video priors.

Therefore, existing EVFI methods primarily focus on interpolating intermediate frames and remain limited to modeling 2D appearance. In contrast, our work directly interpolates 3D geometry in continuous time, recovering both scene structure and camera motion at arbitrary time instants.

2.3 Event-aided/-based 3D Models

Beyond frame-based 3D geometry estimation, several recent works explore event cameras either as a complementary sensing modality or as a standalone input for robust 3D perception. Thanks to their high temporal resolution and high dynamic range, event cameras provide rich motion cues that can benefit geometric estimation under fast motion and challenging imaging conditions.

ADAE (Peng et al., 2026) fuses events with monocular depth models (Yang et al., 2024a;b) to improve robustness under extreme illumination and motion blur, while retaining per-frame depth prediction. EAG3R (Wu et al., 2025) augments MonST3R with event-derived motion cues for robust depth and pose estimation under extreme lighting and fast motion, with predictions tied to discrete frame timestamps. EventVGGT (Ren et al., 2026) distills VGGT's geometric priors into event representations for temporally consistent depth estimation. Its predictions correspond to fixed temporal windows rather than arbitrary continuous query times. Event3R (Huang et al., 2026) explores event-only 3D reconstruction by aggregating spatio-temporal event features, but its geometric predictions are still tied to a fixed reference time rather than arbitrary continuous-time queries.

In summary, existing event-aided and event-based 3D models show that event cameras provide valuable motion information for 3D geometry estimation. However, they either enhance geometric predictions at discrete frame timestamps or aggregate events in a temporal window to predict geometry at a fixed reference time. None of them explicitly models geometry in continuous-time space or recovers depth and ego-motion at arbitrary timestamps between the input frames. In contrast, our work explicitly leverages the quasi-continuous event stream together with frame-based pointmaps to interpolate geometric representations at arbitrary query times, enabling continuous-time depth and camera pose estimation for the first time (see Table 4 in the Appendix for a detailed comparison).

3 Method

3.1 Preliminaries

Our model achieves continuous-time estimation of depth and camera poses by interpolating pointmaps from frame-based models like DUSt3R. Let us begin with a brief review of DUSt3R, as well as its follow-ups MonST3R and Align3R, which extend it from static to dynamic scenes.

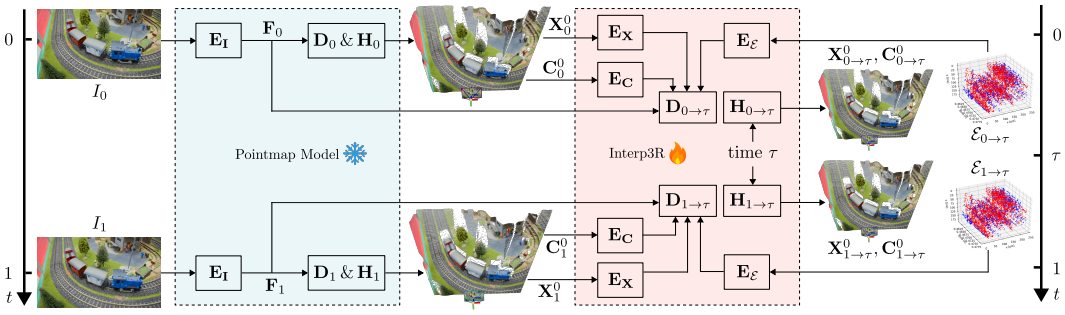

Figure 2: Pipeline for pointmap prediction and interpolation. The pointmap model first takes as input two frames and outputs the source pointmaps. Interp3R then interpolates the pointmap at the arbitrary target time $\tau \in (0, 1)$ from two directions, with the first part ($0 \to \tau$, "forward") and the second part ($1 \to \tau$, "backward") event data.

Pointmap-based Geometry Estimation. Recent advances in pose-free 3D reconstruction adopt pointmap representations to directly regress dense geometry from image pairs. DUSt3R predicts per-pixel 3D coordinates (pointmaps) for an image pair (I_0, I_1) using a transformer-based model:

$$\left(\mathbf{X}_0^0, \mathbf{X}_1^0\right) = \mathcal{F}_{\text{DUSt3R}}(I_0, I_1), \tag{1}$$

where each pointmap assigns a 3D point to every pixel in the reference view. $\mathbf{X}_0^0$ and $\mathbf{X}_1^0$ indicate the pointmaps of I_0 and I_1 respectively, and both are expressed in the camera coordinate of I_0. In general, $\mathbf{X}_m^n$ indicates the pointmap of I_m expressed in the camera coordinate of I_n. These pairwise pointmaps encode 3D scene geometry and can be aligned via a global optimization to recover consistent depth and camera poses across multiple views.

Extension to Dynamic Scenes. MonST3R extends DUSt3R to dynamic videos by fine-tuning its pointmap model on dynamic datasets. Specifically, MonST3R finds a global state X^*_{global} by minimizing an objective that combines alignment, motion smoothness, and image-based flow constraints:

$$\mathcal{L}_{\text{MonST3R}}(X^*_{\text{global}}) = \mathcal{L}_{\text{align}} + w_{\text{smooth}}\mathcal{L}_{\text{smooth}} + w_{\text{flow}}\mathcal{L}_{\text{flow}}. \tag{2}$$

This formulation enables joint estimation of video depth and camera motion in dynamic scenes.

Incorporating Monocular Depth Priors. Instead of relying solely on RGB images, Align3R injects per-frame monocular depth estimates to guide pointmap regression, providing stronger geometric initialization and improving depth accuracy and stability, especially for dynamic scenes.

3.2 Pairwise Pointmap Interpolation

As depicted in Figure 2, a pointmap model (we use Align3R as the main model in this work) takes as input an image pair (I_0, I_1) and outputs the pairwise pointmaps $(\mathbf{X}_0^0, \mathbf{X}_1^0)$ as well as the corresponding confidence maps $(\mathbf{C}_0^0, \mathbf{C}_1^0)$ (i.e., a confidence value per 3D point).

Then, Interp3R takes as input these pointmap predictions, the RGB features $(\mathbf{F}_0, \mathbf{F}_1)$ of the two input frames (I_0, I_1) and the bidirectional event data $(\mathcal{E}_{0\to\tau}, \mathcal{E}_{1\to\tau})$, to interpolate the pointmaps $(\mathbf{X}_{0\to\tau}^0, \mathbf{X}_{1\to\tau}^0)$ at the target time $\tau \in (0, 1)$ from both directions $0 \to \tau$ and $1 \to \tau$. Note that both interpolated pointmaps are expressed in the camera coordinate of I_0.

Incorporating Event Data and Pointmaps. The events between the two frames are divided into a forward set, $\mathcal{E}_{0\to\tau}$, and a backward set, $\mathcal{E}_{\tau\to1}$. The backward set $\mathcal{E}_{\tau\to1}$ is formed by reversing the timestamps and polarities of all events, resulting in $\mathcal{E}_{1\to\tau}$. After voxelization into spatiotemporal grids, denoted as $\mathcal{V}_{0\to\tau}$ and $\mathcal{V}_{1\to\tau}$, they are processed by the same event encoder $\mathbf{E}_\mathcal{E}$, as both of them describe motion from the source frame to the target interpolation time. Similarly, the predicted pointmaps $(\mathbf{X}_0^0, \mathbf{X}_1^0)$ and the corresponding confidence $(\mathbf{C}_0^0, \mathbf{C}_1^0)$ are also encoded by the same type of small vision transformer (ViT) encoders ($\mathbf{E_X}$ and $\mathbf{E_C}$) respectively. The extracted features from the pointmap predictions as well as the event voxels are then gradually fused in the Interp3R decoders $(\mathbf{D}_{0\to\tau}, \mathbf{D}_{1\to\tau})$ using zero convolution (Zhang et al., 2023).

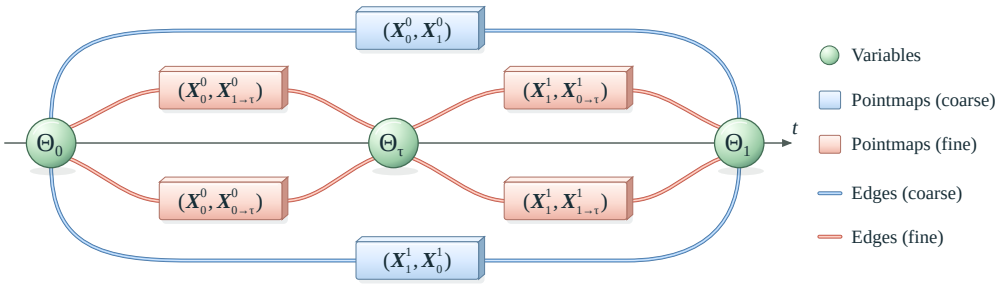

Figure 3: Coarse-to-fine global alignment. Green nodes Θ_t collect the camera intrinsics, pose and depth variables at time t. Each $\mathbf{X}_m^n$ is a per-pixel 3D pointmap of I_m expressed in I_n's camera coordinate. In the coarse alignment (blue), the pointmap pairs connect and constrain only the two source nodes at $t = 0$ and $t = 1$. In the fine alignment (red), the interpolated node at $t = \tau$ is connected to both source nodes through the red pointmap pairs and all node variables are refined.

The decoders and heads of Interp3R have the same structure as those of Align3R (but other choices are possible, as we show compatibility with MonST3R in Section 4.5), and are initialized with Align3R's weights. Before being trained, Interp3R exactly outputs $(\mathbf{X}_0^0, \mathbf{X}_1^0)$, i.e., the pointmaps of the two input frames. As training proceeds, the information about the source pointmaps and about the motion from the source to the target time τ are gradually injected to the Interp3R decoder. In this way, we do not abruptly change Align3R decoders and heads, instead, the predicted pointmaps are progressively transported from the source to the target time, ensuring stable training.

In most EVFI methods (Tulyakov et al., 2021; 2022; Ma et al., 2024), forward and backward events are simply voxelized into spatiotemporal volumes with identical channel dimensions and directly fed into the network for motion prediction. The interpolation time τ is only implicitly encoded through temporal truncation of events, without explicit modeling. As a result, the interpolation quality becomes highly sensitive to the interpolation time, especially in the presence of non-uniform motion or complex scene dynamics. Inspired by Jin et al. (2025), we explicitly incorporate a time encoding into the prediction heads, conditioning the model on the target time $\tau \in (0, 1)$.

Training Objective. As shown in Figure 2, we only train the Interp3R part (i.e., the Align3R component is frozen), and use the confidence-aware scale-invariant 3D regression loss of DUSt3R. We normalize the interpolated $\mathbf{X}_\tau$ and groundtruth pointmaps $\bar{\mathbf{X}}_\tau$ at $t = \tau$ using the scale factors calculated from the Align3R-predicted and groundtruth pointmaps at $t = 0$ and 1: $z = \text{norm}(\mathbf{X}_0, \mathbf{X}_1)$ and $\bar{z} = \text{norm}(\bar{\mathbf{X}}_0, \bar{\mathbf{X}}_1)$[1], respectively. The final loss for image pair (I_0, I_1) is:

$$\mathcal{L}_{\text{interp}} = \sum_{v\in\{0,1\}} \mathbf{C}_{v\to\tau} \left\| \frac{1}{z}\mathbf{X}_{v\to\tau} - \frac{1}{\bar{z}}\bar{\mathbf{X}}_{v\to\tau} \right\| - \alpha \log \mathbf{C}_{v\to\tau} \tag{3}$$

where $\mathbf{C}_{v\to\tau}$ is the pixel-wise confidence of $\mathbf{X}_{v\to\tau}$, and α is a hyper-parameter that controls the regularization term (Wang et al., 2024). This loss function is the sum of interpolation errors (3D weighted Euclidean distance) from both temporal directions.

3.3 Coarse-to-Fine Global Alignment

After obtaining the pointmaps for all the frame and interpolation timestamps, we perform a coarse-to-fine global alignment to solve for the expected camera intrinsics, poses and depth at all those timestamps. We illustrate the proposed coarse-to-fine alignment in the case of an image pair in Figure 3, where Θ_0, Θ_τ and Θ_1 are the desired variables. During coarse alignment, we align the source pointmaps produced by Align3R (in blue boxes) to solve for Θ_0 and Θ_1, which is the same as the global alignment in Align3R. Note that dynamic scenes can be handled due to the capability of Align3R. The objective function of the coarse alignment is (2), i.e., the one in MonST3R, where RAFT (Teed & Deng, 2020) is used to provide the optical flow required by the flow loss.

The fine alignment stage has two steps: (i) fixing the variables optimized in the coarse stage, aligning the interpolated pointmaps to the source pointmaps to solve for the variables at the target time, Θ_τ;

[1] where a bar denotes groundtruth and *norm* computes the mean distance from the points to the world origin.

(ii) finetuning all variables (Θ_0, Θ_τ, Θ_1) to further improve consistency. In the second step, we add a scale-invariant coarse-translation prior term $\mathcal{L}_{\text{trans}}$ to (2) to preserve the shape of the coarse-frame trajectory while allowing one global scale degree of freedom, whose formulation is described in (7) in the supplementary. Therefore, the final loss of the second step is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MonST3R}} + w_{\text{trans}}\mathcal{L}_{\text{trans}}$. In the fine alignment, the flow loss uses optical flow from E-RAFT (Gehrig et al., 2021b). In practice, the target time τ can be set to any value between 0 and 1, and an arbitrary number of interpolated pointmaps can be fine-aligned, leading to continuous-time depth and pose estimation.

Following DUSt3R, we symmetrize every consecutive image pair (I_0, I_1) and event sets $(\mathcal{E}_{0\to\tau}, \mathcal{E}_{1\to\tau})$ to obtain a new pair (I_1, I_0) with $(\mathcal{E}_{1\to\tau}, \mathcal{E}_{0\to\tau})$. We also use this new pair to interpolate the pointmaps $(\mathbf{X}^1_{1\to\tau}, \mathbf{X}^1_{0\to\tau})$, where the target time is $1-\tau$ and the output pointmaps are expressed in the camera coordinate of I_1. In this way, we obtain four pointmaps with confidence at the target position, which are inferred from two directions while expressed in two different camera coordinates. These four measurements are then naturally fused according to pixel-wise confidence to obtain the estimation of depth and pose at $t = \tau$ in the fine alignment stage.

4 Experiments

4.1 Experimental Setup

Implementation details. We train our model on six NVIDIA RTX A6000 GPUs with a total batch size of 12 for 50 epochs. We use the AdamW optimizer with a learning rate of 10^{-4}. Training is performed solely on the PointOdyssey (Zheng et al., 2023) dataset. Each training sample consists of three images corresponding to timestamps $t = 0$, $t = \tau$, and $t = 1$ (i.e., a triplet). The temporal span of each triplet ranges from 2 to 10 frames. We also symmetrize the triplets for training to encourage the consistency of the trained model, as DUSt3R does. Before training, we precompute all valid triplets and construct a uniformly distributed sampling pool by balancing both the left-right temporal spans and the relative position of the middle frame. This ensures that the model observes interpolation cases with different spans and relative positions evenly. During training, each epoch randomly samples 20,000 triplets from this pool to train the model.

Baselines. Similar to the evaluation of EVFI (Tulyakov et al., 2021), we skip $k = \{1, 3, 7, 15\}$ frames, feed the remaining frames as well as the event data in between to Interp3R to solve for depth and camera poses. These skips enable quantitative evaluation at groundtruth timestamps (see supplementary Figure 7 for arbitrary-time qualitative results, e.g., $t = \{1/\pi, \sqrt{10}/4\}$). Since our method is the first to address continuous-time depth and pose estimation tasks with events and frames, we create baselines based on concatenating state-of-the-art individual-task methods: we first leverage VFI methods: frame-based RIFE (Huang et al., 2022), and event-based Timelens (Tulyakov et al., 2021) as well as VDM-EVFI (Chen et al., 2025a) to interpolate skipped frames, and then input all frames to Align3R to solve for depth and poses.

4.2 Continuous-time Depth Estimation

Evaluation datasets. We evaluate our approach on both synthetic and real-world benchmarks. The real datasets include Bonn (Palazzolo et al., 2019) and DSEC (Gehrig et al., 2021a), while the synthetic datasets include Sintel (Butler et al., 2012) and the validation split of PointOdyssey (Zheng et al., 2023). Since Bonn, Sintel, and PointOdyssey do not contain event data, we generate events for them using vid2e (Gehrig et al., 2020). For DSEC, we use recorded real events and spatially align RGB frames to them, then evaluate our method with real frame and event measurements.

Evaluation metrics. Following prior works (Zhang et al., 2025; Lu et al., 2025), we first align the predicted depth maps to the groundtruth (GT) depth using a single scale and shift before computing the evaluation metrics. To make the comparison of temporal consistency fair across different methods, the scale and shift are estimated once for the entire image sequence rather than independently for each frame. We report two standard metrics: the absolute relative error (Abs Rel $\downarrow$) and the fraction of inlier pixels with $\delta < 1.25$ ($\uparrow$). To highlight the performance on the interpolated frames, the metrics are calculated on the skipped frames.

Comparison with baseline methods. Table 1 shows that Interp3R achieves the lowest Abs Rel on all four datasets at every skip value, together with the highest inlier ratio on PointOdyssey, Sintel,

Table 1: Depth estimation results. **Best** and second-best are highlighted.

Dataset	Method	skip = 1		skip = 3		skip = 7		skip = 15	
		Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑
PointOdyssey	RIFE	0.097	0.925	0.106	0.911	0.146	0.891	0.196	0.846
	TimeLens	0.087	0.925	0.099	0.912	0.142	0.858	0.215	0.777
	VDM-EVFI	0.108	0.889	0.113	0.875	0.118	0.861	0.137	0.839
	Interp3R (Ours)	**0.061**	**0.954**	**0.061**	**0.957**	**0.072**	**0.943**	**0.105**	**0.886**
Sintel	RIFE	0.381	0.538	0.478	0.532	0.515	0.522	0.580	0.507
	TimeLens	0.377	0.548	0.441	0.545	0.530	0.496	0.595	0.453
	VDM-EVFI	0.366	0.542	0.390	0.553	0.397	0.559	0.424	0.543
	Interp3R (Ours)	**0.334**	**0.610**	**0.328**	**0.621**	**0.354**	**0.613**	**0.417**	**0.573**
Bonn	RIFE	0.075	0.956	0.079	0.952	0.098	0.932	0.157	0.886
	TimeLens	0.077	0.956	0.082	0.949	0.104	0.922	0.226	0.758
	VDM-EVFI	0.074	**0.959**	0.076	**0.955**	0.076	0.956	0.076	**0.956**
	Interp3R (Ours)	**0.066**	0.949	**0.064**	0.952	**0.064**	**0.959**	**0.072**	0.955
DSEC	RIFE	0.117	0.860	0.131	0.839	0.154	0.806	0.179	0.790
	TimeLens	0.127	0.837	0.151	0.789	0.203	0.694	0.311	0.508
	VDM-EVFI	0.150	0.786	0.153	0.780	0.166	0.744	0.185	0.705
	Interp3R (Ours)	**0.099**	**0.897**	**0.095**	**0.910**	**0.105**	**0.883**	**0.127**	**0.858**

Table 2: Pose evaluation results. **Best** and second-best are highlighted.

Dataset	Method	skip = 1			skip = 3			skip = 7			skip = 15		
		ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓
Sintel	RIFE	0.301	0.104	1.259	**0.292**	**0.115**	1.290	0.328	0.146	1.675	0.434	0.157	2.547
	TimeLens	0.321	0.114	1.265	0.385	0.174	1.412	0.566	0.175	2.410	0.599	0.141	3.719
	VDM-EVFI	0.319	0.127	1.314	0.339	0.128	1.193	0.405	0.148	1.351	0.379	**0.118**	**1.496**
	Interp3R (Ours)	**0.214**	**0.089**	**0.815**	**0.292**	0.140	**0.908**	**0.219**	**0.101**	**1.214**	**0.279**	0.154	1.869
Bonn	RIFE	0.011	0.007	0.749	0.012	0.007	0.746	0.013	**0.007**	0.827	0.015	0.008	0.782
	TimeLens	0.011	0.008	0.762	0.012	0.008	0.765	0.015	0.009	0.769	0.017	0.010	1.482
	VDM-EVFI	0.016	0.011	0.816	0.017	0.011	0.792	0.016	0.010	0.772	0.019	0.010	0.835
	Interp3R (Ours)	**0.006**	**0.006**	**0.500**	**0.006**	**0.007**	**0.508**	**0.009**	**0.007**	**0.504**	**0.008**	**0.006**	**0.514**
ECD	RIFE	0.175	0.065	3.144	0.178	0.083	3.671	0.197	0.104	3.640	0.234	0.104	4.082
	TimeLens	0.184	0.065	3.184	0.262	0.098	3.693	0.339	0.105	3.643	0.359	0.113	3.773
	VDM-EVFI	0.228	0.084	3.725	0.214	0.084	3.644	0.226	0.093	3.493	0.223	0.083	3.380
	Interp3R (Ours)	**0.121**	**0.027**	**2.213**	**0.138**	**0.039**	**2.346**	**0.172**	**0.048**	**2.512**	**0.218**	**0.071**	**2.478**
EDS	RIFE	0.045	0.023	0.487	0.046	0.023	0.533	**0.040**	0.023	0.894	0.091	0.053	1.684
	TimeLens	0.047	0.025	0.529	0.051	0.026	0.510	0.070	0.037	0.700	0.233	0.048	1.776
	VDM-EVFI	0.103	0.050	1.164	0.104	0.046	1.058	0.112	0.040	1.016	0.106	**0.047**	1.026
	Interp3R (Ours)	**0.024**	**0.006**	**0.346**	**0.023**	**0.007**	**0.339**	**0.040**	**0.017**	**0.346**	**0.086**	0.049	**0.659**

and DSEC. On PointOdyssey, it reduces Abs Rel over the strongest baseline by approximately 23%–39%, while on Bonn its accuracy remains nearly constant as the frame gap grows.

DSEC provides a particularly relevant test of generalization because it contains real images and event measurements from outdoor driving scenes, whereas training uses only synthetic PointOdyssey data. Interp3R leads both depth metrics across all temporal intervals, reducing Abs Rel by approximately 15%–32% relative to the strongest baseline. The improvement remains substantial under sparse observations: at skips 3, 7, and 15, the error reduction is approximately 27%–32%. At skips 7 and 15, the inlier ratio also improves by approximately 9%–10% relative to the best baseline, indicating that the gains extend to a larger fraction of accurately reconstructed pixels.

As the frame gap increases, RIFE and TimeLens degrade markedly on DSEC, while Interp3R retains lower error throughout. This suggests that events provide useful motion cues for recovering intermediate geometry, especially when image observations are far apart. Since TimeLens and VDM-EVFI also use events, the gains suggest that the benefit extends beyond simply adding the sensing modality. Interp3R couples event motion cues with source pointmaps during 3D prediction, allowing intermediate depth to be estimated directly within the geometric representation.

The results also show that these cues remain effective with real sensor data, supporting the transfer of our frame-event fusion from synthetic training to real-world continuous-time depth estimation.

4.3 Continuous-time Pose Estimation

Evaluation datasets. For camera pose evaluation, we largely follow the protocol used in Align3R and its preceding works. The evaluation is conducted on four datasets: the real-world Bonn dataset,

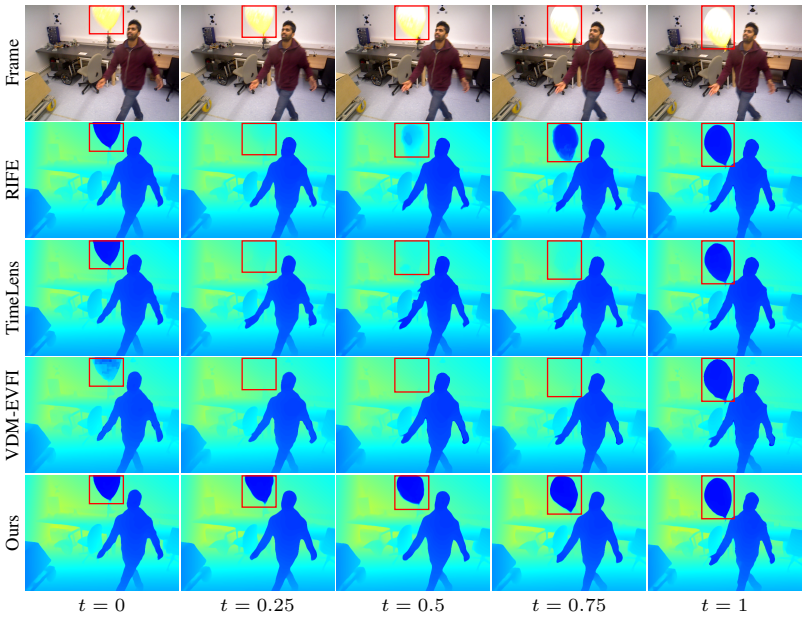

Figure 4: Qualitative comparison of depth estimation on the Bonn dataset (skip = 3).

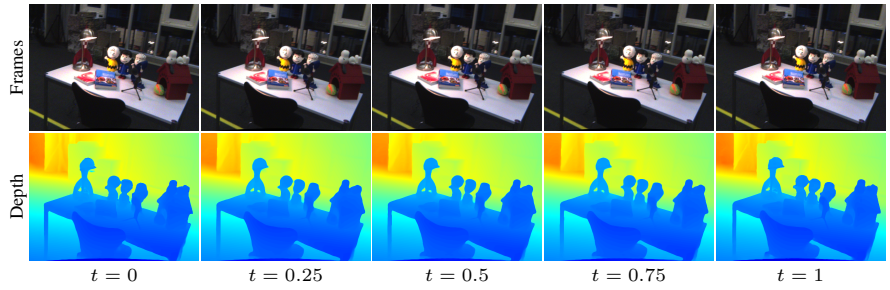

Figure 5: Qualitative estimated depth results on the EDS dataset (skip = 3).

the synthetic Sintel dataset, and the real event camera datasets ECD (Mueggler et al., 2017) and EDS (Hidalgo-Carrió et al., 2022). For ECD and EDS, we directly use the recorded event measurements, allowing us to evaluate continuous-time pose estimation with real event data.

Evaluation metrics. Following prior studies (Zhang et al., 2025; Lu et al., 2025), we report three commonly used metrics: ATE↓ (Absolute Trajectory Error [m]), RTE↓ (Relative Translation Error [m]), and RRE↓ (Relative Rotation Error [°]). ATE measures the discrepancy between the estimated trajectory and the ground-truth trajectory. RTE and RRE quantify the average local translation and rotation errors, respectively, computed between consecutive poses.

Comparison with baseline methods. Table 2 shows strong improvements in pose estimation across evaluated datasets. Interp3R reports the best ATE on all datasets at every skip, and best RTE and RRE in 13/16 and 15/16 settings, respectively. The advantages transfer to real event camera data on ECD and EDS, despite training exclusively on synthetic sequences.

On ECD, Interp3R achieves the best ATE, RTE, and RRE at every skip value. Relative to the strongest baseline for each metric, it reduces RRE by 27%–36% and RTE by 14%–58% across the four skips. The consistent improvements in both local motion and global trajectory accuracy indicate that event cues benefit the recovered camera poses as a whole. At skip $= 7$, the reductions in ATE, RTE, and RRE are 13%, 48%, and 28%, respectively, showing that the advantage persists when the input frames are sparsely sampled. On EDS, Interp3R achieves the lowest ATE and RRE at every skip value, with particularly strong gains in rotation estimation: RRE decreases by 29%–50% relative to the strongest baseline, and the largest improvement occurs at skip $= 7$. At skips 1 and 3, ATE is reduced by 46%–51%, while RTE drops by about 71%. All these results show that the proposed method achieves precise pose estimation on real-world frames and events.

Table 3: **Ablation studies.** Depth estimation on DSEC and camera pose estimation on EDS.

Method	Depth estimation (DSEC)						Camera pose estimation (EDS)								
	skip = 1		skip = 3		skip = 7		skip = 1			skip = 3			skip = 7		
	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	$\delta < 1.25$ ↑	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓
w/o events	0.103	0.890	0.108	0.893	0.123	0.856	0.028	0.010	0.454	0.033	0.015	0.441	0.090	0.056	0.466
MonST3R + Interp3R	0.122	0.855	0.111	0.880	0.118	**0.887**	**0.024**	**0.005**	**0.337**	0.031	0.008	0.373	0.061	0.033	0.401
Align3R + Interp3R	**0.099**	**0.897**	**0.095**	**0.910**	**0.105**	0.883	**0.024**	0.006	0.346	**0.023**	**0.007**	**0.339**	**0.040**	**0.017**	**0.346**



(a) RIFE (b) TimeLens (c) VDM-EVFI (d) Ours

Figure 6: 3D reconstruction of a driving sequence fro m the DSEC dataset (skip = 3).

4.4 Qualitative Results

A qualitative comparison of estimated depth maps on Bonn is presented in Figure 4. All baseline methods fail to estimate reasonable depth for the balloon. More specifically, the balloon completely disappears in the depth maps of TimeLens and VDM-EVFI while partially shows up in those of RIFE. For VDM-EVFI, the error is further propagated to the source frame ($t = 0$). This implies that the inaccuracy of the 2D VFI results is significantly amplified in the 3D geometry estimation, which leads to unreasonable and temporally inconsistent depth values. In contrast, our method produces spatially smooth and temporally consistent depth estimation.

Figure 5 shows qualitative results on the EDS dataset, estimated from real images and events. The depth maps at $t = \{0.25, 0.5, 0.75\}$ recover the scene's fine geometric structures very well. This shows our synthetically-trained model generalizes well to real images and events. Finally, Figure 6 depicts that our method retains a clearer building facade and a more coherent road surface than the baselines. The camera poses (colored frusta) remain organized along the street, supporting joint reconstruction of scene geometry from real events. More results are given in the Appendix.

4.5 Ablation Studies

Table 3 evaluates event input and backbone compatibility on DSEC depth and EDS pose estimation.

Effects of Event Data. We remove the event input and encoder $\mathbf{E}_{\mathcal{E}}$ while keeping other training settings unchanged. Events improve every reported depth and pose metric, with the Abs Rel reduction growing from 4% to 15% as the skip increases from 1 to 7, which shows event data becomes more important as the frame skip grows. At skip $= 7$, events reduce ATE and RTE by 55% and 70%, respectively, demonstrating their importance for recovering motion between sparse frames.

Compatibility with MonST3R. We replace Align3R with MonST3R at inference time, retaining the Interp3R model trained on Align3R without finetuning. The resulting variant remains competitive across the evaluated skips and reduces RTE by 17% at skip $= 1$ relative to Align3R + Interp3R. This transfer demonstrates that Interp3R can serve as a plug-and-play module for extending DUSt3R-like frame-based pointmap models to continuous-time estimation of camera motion and scene geometry.

5 Conclusion

This paper introduced Interp3R, a novel framework that extends frame-based pointmap models to continuous-time 3D geometry estimation (depth maps and camera poses) by leveraging, for the first time, co-located asynchronous event camera data. Thorough experiments showed that our approach consistently outperforms state-of-the-art baselines that rely on two-stage pipelines of video frame interpolation followed by frame-based geometry estimation. We hope our work highlights the potential of combining event-based sensing with modern 3D visual foundation models and opens new directions for temporally continuous 3D perception in dynamic environments.

AI DISCLOSURE

In this work, we used generative AI tools to (*tasks with required disclosure*): assist with translation clean and reformat dataset, improve readability of some sentences in the text after an initial manual writing, write LaTeX macros to facilitate the layout of multiple images in the same figure.

We have not used generative AI tools to (*other tasks with required disclosure*): generate synthetic data sets, design or provide feedback on research methodology or experiments, help develop theoretical models or conceptual frameworks, support qualitative and thematic data analysis, interpret results, implement methods, come up with the main ideas of the paper (contributions 1-3), generate the quantitative or qualitative results in any of the tables or figures in the experiments (all of our results are reproducible with our code, which will be open-sourced upon acceptance), hallucinate references.

The rest of the *required disclosure tasks*: formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, are not applicable to this work.

Additionally, we used generative AI tools to (*tasks with recommended disclosure*): edit a research paper to improve readability, improving the diagram in Fig 3 from an initial version of it, create or edit software code, sourcing/searching for information.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240x180 130dB 3μs latency global shutter spatiotemporal vision sensor. *IEEE J. Solid-State Circuits*, 49 (10):2333–2341, 2014. doi: 10.1109/JSSC.2014.2342715.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Eur. Conf. Comput. Vis. (ECCV)*, pp. 611–625, 2012. doi: 10.1007/978-3-642-33783-3_44.

Jingxi Chen, Brandon Y Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermuller, Christopher A Metzler, and Yiannis Aloimonos. Repurposing pre-trained video diffusion models for event-based video interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 12456–12466, June 2025a. doi: 10.1109/cvpr52734.2025.01162.

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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 22831–22840, 2025b. doi: 10.1109/CVPR52734.2025.02126.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to Events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 3583–3592, 2020. doi: 10.1109/CVPR42600.2020.00364.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robot. Autom. Lett.*, 6(3):4947–4954, 2021a. doi: 10.1109/LRA.2021.3068942.

Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *Int. Conf. 3D Vision (3DV)*, pp. 197–206, 2021b. doi: 10.1109/3DV53792.2021.00030.

Shuang Guo and Guillermo Gallego. CMax-SLAM: Event-based rotational-motion bundle adjustment and SLAM system using contrast maximization. *IEEE Trans. Robot.*, 40:2442–2461, 2024. doi: 10.1109/TRO.2024.3378443.

Shuang Guo and Guillermo Gallego. Event-based photometric bundle adjustment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):9280–9297, 2025. doi: 10.1109/TPAMI.2025.3586497.

Shuang Guo, Friedhelm Hamann, and Guillermo Gallego. Unsupervised joint learning of optical flow and intensity with event cameras. In *Int. Conf. Comput. Vis. (ICCV)*, pp. 7980–7989, 2025. doi: 10.48550/arXiv.2503.17262.

Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 5781–5790, June 2022. doi: 10.1109/CVPR52688.2022.00569.

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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2005–2015, 2025. doi: 10.1109/CVPR52734.2025.00193.

Jian Huang, Haotian Shen, Xinhao Lou, Chengrui Dong, Wenpu Li, and Peidong Liu. Event3r: Asynchronous-to-global 3d reconstruction from event camera via spatial-temporal feature aggregation. *arXiv preprint arXiv:2607.15727*, 2026.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Eur. Conf. Comput. Vis. (ECCV)*, pp. 624—-642, 2022. doi: 10.1007/978-3-031-19781-9_36.

Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4D: Learning how things move in 3D from internet stereo videos. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 10497–10509, 2025. doi: 10.1109/CVPR52734.2025.00982.

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 18032–18042, June 2023. doi: 10.1109/CVPR52729.2023.01729.

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

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3R: Aligned monocular depth estimation for dynamic videos. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 22820–22830, 2025. doi: 10.1109/CVPR52734.2025.02125.

Yongrui Ma, Shi Guo, Yutian Chen, Tianfan Xue, and Jinwei Gu. TimeLens-XL: Real-time event-based video frame interpolation with large motion. In *Eur. Conf. Comput. Vis. (ECCV)*, pp. 178—-194, October 2024. doi: 10.1007/978-3-031-72907-2_11.

Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *Int. J. Robot. Research*, 36(2):142–149, 2017. doi: 10.1177/0278364917691115.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pp. 7855–7862, 2019. doi: 10.1109/IROS40897.2019.8967590.

Shihan Peng, Yuyang Xiong, Hanyu Zhou, Zhiwei Shi, Haoyue Liu, Gang Chen, Luxin Yan, and Yi Chang. Adapting depth anything to adverse imaging conditions with events. *arXiv e-prints*, 2026.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 724–732, 2016. doi: 10.1109/CVPR.2016.85.

Yinrui Ren, Jinjing Zhu, Kanghao Chen, Zhuoxiao Li, Jing Ou, Zidong Cao, Tongyan Hua, Peilun Shi, Yingchun Fu, Wufan Zhao, and Hui Xiong. Eventvggt: Exploring cross-modal distillation for consistent event-based depth estimation. *arXiv preprint arXiv:2603.09385*, 2026.

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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 22841–22852, 2025. doi: 10.1109/CVPR52734.2025.02127.

Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 18043–18052, 2023. doi: 10.1109/CVPR52729.2023.01730.

Zachary Teed and Jia Deng. RAFT: Recurrent all pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis. (ECCV)*, pp. 402–419, 2020. doi: 10.1007/978-3-030-58536-5_24.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 16150–16159, June 2021. doi: 10.1109/CVPR46437.2021.01589.

Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 17734–17743, June 2022. doi: 10.1109/CVPR52688.2022.01723.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 20697–20709, 2024. doi: 10.1109/CVPR52733.2024.01956.

Xiaoshan Wu, Yifei Yu, Xiaoyang Lyu, Yihua Huang, Bo Wang, Baoheng Zhang, Zhongrui Wang, and Xiaojuan Qi. EAG3R: Event-augmented 3D geometry estimation for dynamic and extreme-lighting scenes. *arXiv e-prints*, 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 10371–10381, 2024a. doi: 10.1109/CVPR52733.2024.00987.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024b.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *Int. Conf. Learn. Representations (ICLR)*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis. (ICCV)*, pp. 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.

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 *Int. Conf. Comput. Vis. (ICCV)*, pp. 19798–19808, 2023. doi: 10.1109/ICCV51070.2023.01818.