Depth AnyEvent: A Cross-Modal Distillation Paradigm for Event-Based Monocular Depth Estimation

Luca Bartolomei*,†, Enrico Mannocci†, Fabio Tosi†, Matteo Poggi*,†, Stefano Mattoccia*,†

University of Bologna, †Department of Computer Science and Engineering,
*Advanced Research Center on Electronic System (ARCES)

Project Page

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Introduction

Event cameras offer microsecond temporal resolution and robustness to challenging conditions, but suffer from limited training data for monocular depth estimation.

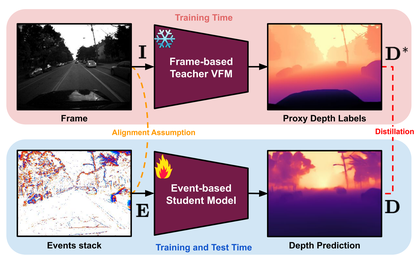
Key Challenges:

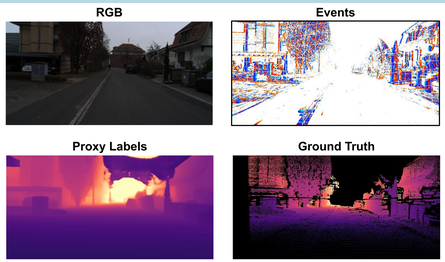
- **Data Scarcity.** Lack of large-scale event datasets with dense depth annotations for training.
- **Sparse Information.** Events only trigger at texture boundaries and moving objects.
- **Domain Gap.** Limited transferability from image-based models to event representations.

Key Contributions:

- **Cross-Modal Distillation:** Leverage VFMs to generate dense proxy labels effectively.
- **Robustness Improvement:** Cast image-based models to event domain through fine-tuning.
- **Unified Approach:** Novel DepthAnyEvent-R with temporal modeling via ConvLSTM.
- **State-of-the-Art:** Superior performance on MVSEC & DSEC benchmark datasets.

Cross-Modal Distillation Strategy

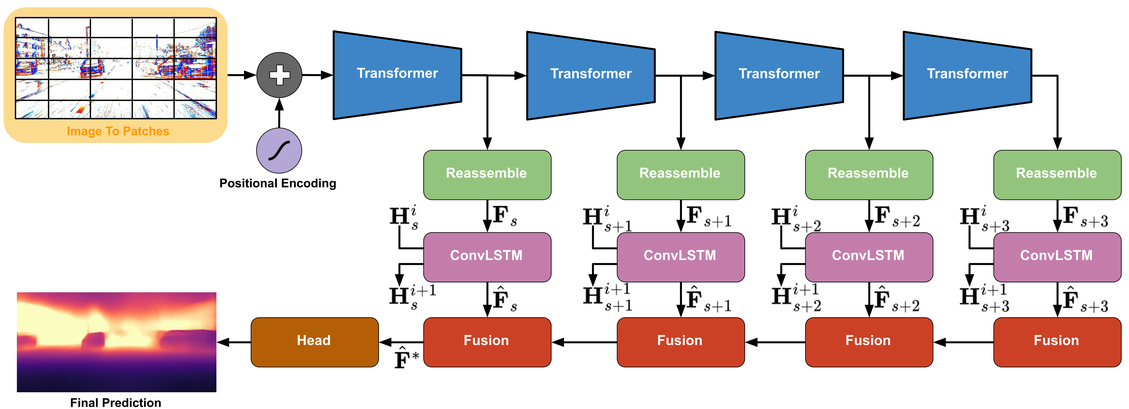




- **Teacher VFM on RGB Frames.** A pre-trained Visual Foundation Model (DAv2 ViT-Large) processes spatially aligned input frames to generate dense proxy depth labels.
- **Event-based Student Model.** The student takes aligned event stacks as input and predicts a depth map.
- **Cross-Modal Distillation.** Knowledge is effectively transferred from the RGB teacher to the event-based student model, enabling robust training without large-scale annotated event datasets.

Proposed Architectures

- **Event Representation with Tencode.** Encode spatio-temporal event slices into a 3-channel representation, reducing the gap with RGB frames and allowing minimal changes to pre-trained VFMs.
- **Adapting VFMs to Events.** Fine-tune a pre-trained DAv2 ViT-S backbone on event data using the Tencode representation strategy, yielding our adapted model *DepthAnyEvent*.
- **Recurrent VFM for Events.** Extend DepthAnyEvent with multi-scale ConvLSTM modules to effectively exploit temporal cues across sequential event stacks, producing the enhanced recurrent model *DepthAnyEvent-R*.



Experimental Results

Zero-Shot Generalization: Trained on EventScape only, tested on MVSEC and DSEC without any fine-tuning.

Method	MVSEC				DSEC			
	AbsRel ↓	SqRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	SqRel ↓	RMSE ↓	δ_1 ↑
E2Depth	0.527	1.122	7.894	0.363	0.395	0.334	13.258	0.409
EReFormer	0.518	1.012	8.423	0.361	0.297	0.195	11.608	0.524
DepthAnyEvent	**0.466**	0.976	**7.824**	0.408	0.297	0.186	11.072	0.519
DepthAnyEvent-R	0.469	**0.946**	8.064	**0.428**	**0.276**	**0.165**	**10.942**	**0.555**

Fine-tuned Performance. Trained on EventScape and then further fine-tuned on MVSEC and DSEC datasets separately.

Method	MVSEC				DSEC			
	AbsRel ↓	SqRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	SqRel ↓	RMSE ↓	δ_1 ↑
E2Depth	0.420	0.806	7.268	0.432	0.253	0.130	10.119	0.574
EReFormer	0.511	1.057	8.373	0.391	0.286	0.208	11.369	0.569
DepthAnyEvent	0.373	0.715	6.627	0.471	0.201	0.079	8.880	0.664
DepthAnyEvent-R	**0.365**	**0.691**	**6.465**	**0.489**	**0.191**	**0.070**	**8.618**	**0.691**

Supervised vs Distilled Models: Trained on EventScape and then fine-tuned on MVSEC and DSEC datasets separately, either through distillation or on ground-truth depth labels.

Method	MVSEC				DSEC			
	AbsRel ↓	SqRel ↓	RMSE ↓	δ_1 ↑	AbsRel ↓	SqRel ↓	RMSE ↓	δ_1 ↑
DepthAnyEvent Synth	0.466	0.976	7.824	0.408	0.297	0.186	11.072	0.519
DepthAnyEvent Distilled	0.397	0.771	6.910	0.461	0.213	0.095	8.930	0.662
DepthAnyEvent Supervised	**0.373**	**0.715**	**6.627**	**0.471**	**0.201**	**0.079**	**8.880**	**0.664**
DepthAnyEvent-R Synth	0.469	0.946	8.064	0.428	0.276	0.165	10.942	0.555
DepthAnyEvent-R Distilled	0.399	0.781	6.830	0.462	0.226	0.111	9.310	0.638
DepthAnyEvent-R Supervised	**0.365**	**0.691**	**6.465**	**0.489**	**0.191**	**0.070**	**8.618**	**0.691**

Qualitative Results