This paper has been accepted for publication at the
IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, 2025. ©IEEE

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

Luca Bartolomei[*,†] Enrico Mannocci[†] Fabio Tosi[†] Matteo Poggi[*,†] Stefano Mattoccia[*,†]
[*]Advanced Research Center on Electronic System (ARCES)
[†]Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{luca.bartolomei5, fabio.tosi5, m.poggi, stefano.mattoccia}@unibo.it
https://bartn8.github.io/depthanyevent

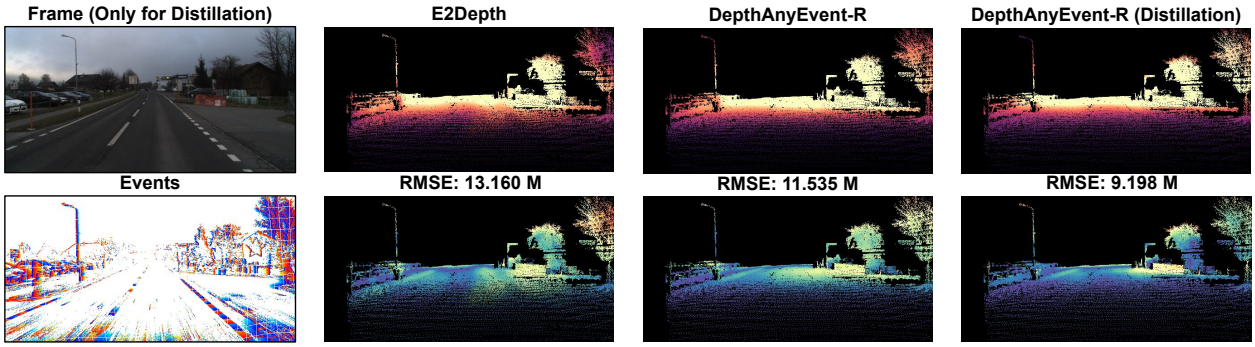

Figure 1. **DepthAnyEvent-R in action.** The first column shows the input frame (used only for distillation) and the corresponding event visualization. The other three columns present depth estimation results from different approaches: E2Depth [15], our DepthAnyEvent-R, and our DepthAnyEvent-R trained with our distillation approach. The top row shows the estimated depth maps while the bottom row depicts their corresponding RMSE visualizations.

Abstract

Event cameras capture sparse, high-temporal-resolution visual information, making them particularly suitable for challenging environments with high-speed motion and strongly varying lighting conditions. However, the lack of large datasets with dense ground-truth depth annotations hinders learning-based monocular depth estimation from event data. To address this limitation, we propose a cross-modal distillation paradigm to generate dense proxy labels leveraging a Vision Foundation Model (VFM). Our strategy requires an event stream spatially aligned with RGB frames, a simple setup even available off-the-shelf, and exploits the robustness of large-scale VFMs. Additionally, we propose to adapt VFMs, either a vanilla one like Depth Anything v2 (DAv2), or deriving from it a novel recurrent architecture to infer depth from monocular event cameras. We evaluate our approach with synthetic and real-world datasets, demonstrating that i) our cross-modal paradigm achieves competitive performance compared to fully supervised methods without requiring expensive depth annotations, and ii) our VFM-based models achieve state-of-the-art performance.

1. Introduction

Depth perception from cameras is paramount for many application fields, such as those concerning the autonomous navigation of agents in complex scenarios or robotic tasks. In these fields, learning-based methods using conventional cameras have obtained compelling results in the last decade. Moreover, this paradigm enabled inferring depth from a single camera, which brings significant advantages compared to multicamera setups in terms of cost, calibration complexity, and physical constraints. Nonetheless, conventional camera systems struggle to provide a prompt and reliable perception of the sensed environment when dealing with highly dynamic scenes resulting from the fast movement of vehicles, drones, robots or in the presence of challenging illumination conditions such as high contrast scenarios, low light, or rapid lighting changes. These limitations are intrinsic to the conventional camera acquisition technology occurring at discrete periodic intervals and with a lim-

ited dynamic range, causing motion blur, over/under exposure, and potentially missing critical information between frames. In contrast, the intrinsic ability to capture scene changes as soon as they appear – with microsecond temporal resolution – and the much higher dynamic range made event cameras [7] ideal for coping with the challenging application fields mentioned above. Event cameras only register brightness changes at each pixel independently, offering exceptional temporal resolution and robustness to lighting variations. However, these features come at the cost of meager information content compared to conventional cameras. Event cameras provide meaningful cues only for a small subset of the framed image with sufficient texture to trigger events, making depth perception from these devices extremely challenging. Moreover, the lack of large datasets with dense ground truth annotations further exacerbates this inherent difficulty, as collecting precise depth ground truth for event data remains costly and technically demanding.

To tackle these issues in a monocular event camera setup, we propose to leverage the effectiveness of image-based Vision Foundation Models (VFMs) for monocular depth estimation. They have demonstrated remarkable capabilities through extensive pretraining on vast image collections, enabling robust depth prediction even in challenging scenario. As the first contribution, given sequences of aligned images and events, we propose a cross-modal distillation strategy that allows us to obtain dense proxy labels from a VFM to train event-based networks. This approach effectively transfers knowledge from the data-rich image domain to the data-sparse event domain. For our purposes, an off-the-shelf device like a DAVIS Camera [29, 32] that incorporates a conventional global shutter camera and an event-based sensor in the same pixel array would suffice to gather spatially aligned event streams and RGB frames.

Additionally, as the second contribution, we propose to adapt VFMs for event-based monocular depth estimation, either using a vanilla model like Depth Anything v2 (DAv2) or a novel recurrent architecture derived from it. To prove the effectiveness of our proposals, we assess the performance with synthetic and real-world datasets, showing that our cross-modal distillation paradigm allows for achieving competitive performance compared to fully supervised approaches, disregarding the need for expensive depth annotation. Moreover, adapting VFMs for monocular depth estimation according to our two proposals is state-of-the-art, setting new benchmarks for event-based depth estimation.

Figure 1 shows the compelling performance of our proposals, and our contributions can be summarized as follows:

- A novel cross-modal distillation paradigm that leverages the robust proxy labels obtained from image-based VFMs for monocular depth estimation.
- An adapting strategy to cast existing image-based VFMs into the event domain effortlessly.
- A novel recurrent architecture based on an adapted image-based VFM.
- Adapting VFMs to the event domain yields state-of-the-art performance, and our distillation paradigm is competitive against the supervision from depth sensors.

2. Related Work

Image-Based Monocular Depth Estimation. Monocular depth estimation has evolved from traditional approaches [27] to deep learning methods [6, 18]. Self-supervised techniques[12, 13, 38] have emerged to address this challenge of limited ground truth data by recasting depth estimation as an image reconstruction task using stereo images or videos. These approaches have been particularly valuable where dense depth annotations are expensive to obtain. A significant step came with affine-invariant models [25, 26] that estimate depth up to an unknown scale and shift, allowing impressive cross-domain generalization capabilities. MiDaS [26] pioneered this direction by training on diverse large-scale datasets, followed by DPT [25] and more recently, the Depth Anything series [33, 34]. These latter models represent the first generation of Visual Foundation Models for monocular depth estimation, leveraging large-scale pretraining and diverse data sources to achieve unprecedented robustness. The effectiveness of these models lies in their ability to combine knowledge from various domains, including internet photo collections [20, 35], LiDAR from autonomous driving scenarios [10], and RGB-D sensors [23]. Recent advances in VFMs have focused on improving metric accuracy through camera parameter integration [14, 36], leveraging generative approaches like diffusion models [5, 17, 28], and addressing temporal consistency[30]. Furthermore, attention-based architectures and transformer models [37] have shown significant improvements in capturing long-range dependencies crucial for accurate depth. Despite recent advances, applying these methods to event-based cameras is still limited by the lack of large-scale annotated datasets. We tackle this by distilling knowledge from frame-based VFMs, enabling accurate depth estimation without costly event data annotations.

Event-based Monocular Depth Estimation. Event-based depth estimation began with supervised approach using recurrent architectures [8, 15, 21] designed to process the temporal information contained in event streams. Advanced models like [8] further expanded this concept by fusing event and RGB data to exploit their complementary charactetistics. Multimodal fusion techniques have also been explored, combining events with LiDAR to generate dense depth maps [3]. To address the scarcity of labeled event data, self-supervised methods have emerged as promising alternatives. Zhu et al. [40] developed a framework that jointly estimates depth, optical flow, and camera poses using stereo consistency and motion blur minimiza-

tion as training signals. Subsequent work [41] eliminated the need for stereo setups by leveraging pose information from consecutive RGB frames aligned with the event camera, enabling dense depth estimation. Despite these advances, event-based depth estimation still falls short compared to frame-based methods.

3. Preliminaries: Event Depth Estimation

Event cameras measure the logarithmic change in brightness over time, and when it changes over a threshold $\pm C$, the associate pixel at position (x_k, y_k) emits at time t_k an asynchronous signal $e_k = (x_k, y_k, p_k, t_k)$ called *event*. Depending on the sign of this change, the event will have polarity $p_k \in \{-1, 1\}$. Each pixel of the $W \times H$ sensor grid of the event camera can independently emit events at any time, producing an asynchronous stream of events $\mathcal{E} = \{e_k\}_{k=1}^{N}$, where N is the total number of fired events.

Given the event history $\mathcal{E}$, previous event-based dense monocular depth estimation models [8, 15, 21] convert the flow of events into a $\mathbf{E} \in \mathbb{R}^{W \times H \times C}$ structured representation – such as Voxel Grids [40] – since the sparse structure of $\mathcal{E}$ is not suitable for standard CNNs. Intentionally, to estimate a depth map $\mathbf{D} \in \mathbb{R}^{W \times H}$ at a given timestamp t_d, events are retrospectively sampled from the stream $\mathcal{E}$, either within a fixed time window (SBT) – *i.e.*, $\mathcal{E}_{t_d}^{\Delta T} = \{e_k \in \mathcal{E} \mid t_d - \Delta T \leq t_k \leq t_d\}$ – or up to a predefined number K of events (SBN) – *i.e.*, $\mathcal{E}_{t_d}^{K} = \{e_k \in \mathcal{E} \mid d - K \leq k \leq d\}$ – and subsequently *stacked* using different strategies, including:

Voxel Grid [40]: The time interval used for sampling events is divided into B uniform bins, where event polarities are accumulated using linear interpolation within each bin of a $\mathbf{E} \in \mathbb{R}^{W \times H \times B}$ stack.

Image-like [21]: A color-based representation where the R and B channels encode positive and negative polarities, respectively, resulting in an RGB image, *i.e.* a $\mathbf{E} \in \mathbb{R}^{W \times H \times 3}$ stack. Unlike the Voxel Grid representation, it does not retain temporal information.

Tencode [16]. A color image representation in which R and B channels encode positive and negative polarities, with G encoding the timestamp relative to the total time-lapse. It produces an RGB image, *i.e.* a $\mathbf{E} \in \mathbb{R}^{W \times H \times 3}$ stack.

For the sake of space, we report only the event representations relevant to our work, but additional details regarding event representations can be found in [1, 11].

4. Proposed Method

Our first goal is to leverage the knowledge of frame-based monocular depth models like DAv2 extracting pseudo labels to train *any* event-based student depth model – *e.g.*, E2Depth – given aligned intensity frames and event stacks. Figure 2 outlines our cross-modal distillation paradigm.

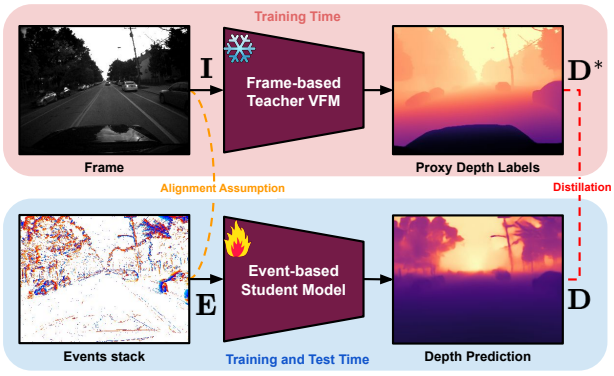

Figure 2. **Proposed Cross-Modal Distillation Strategy.** During training, a VFM teacher processes RGB input frames $\mathbf{I}$ to generate proxy depth labels $\mathbf{D}^*$, which supervise an event-based student model. The student takes aligned event stacks $\mathbf{E}$ as input and predicts the final depth map $\mathbf{D}$.

Moreover, we propose to cast a frame-based model – DAv2 in our experiments – either in its original version or enriching it to exploit temporal cues, to the event domain taking advantage of the massive pre-train performed in the image domain.

4.1. VFMs for Cross-Modal Distillation

Visual Foundation Models have achieved astonishing results mainly due to their peculiar large-scale training procedures. For instance, DAv2 relies on a DINOv2 backbone that was pre-trained with hundreds of millions of images in an unsupervised manner. Furthermore, DAv2 uses tens of millions of pseudo-labeled and millions of labeled images for training. Unfortunately, event data lacks equivalent large-scale datasets [2, 9, 39], substantially precluding comparable training in the event domain. To bridge this gap, we propose leveraging a pre-trained VFM – DAv2 ViT-Large in our experiments– to provide dense supervision for any event-based depth estimation networks, as outlined in Figure 2. During training, a teacher VFM processes a frame, producing the proxy label $\mathbf{D}^*$ (Fig. 3 shows an example) and the student model predicts a depth map $\mathbf{D}$ from the spatially and temporally aligned events. The student model is supervised using a loss $\mathcal{L} = \mathcal{L}_{si} + \lambda \mathcal{L}_{reg}$ composed of a scale-invariant loss $\mathcal{L}_{si}$ and a gradient regularization term $\mathcal{L}_{reg}$ [19]:

$$\mathcal{L}_{si}(\hat{\mathbf{D}}, \hat{\mathbf{D}}^*) = \frac{1}{2\,|\mathbf{M}|} \sum_{(x,y)\in\mathbf{M}} \left(\hat{\mathbf{D}} - \hat{\mathbf{D}}^*\right)^2 \qquad (1)$$

where $\mathbf{M}$ is the set of valid pixels, $\hat{\mathbf{D}} = s\mathbf{D} + t$ and $\hat{\mathbf{D}}^* = \mathbf{D}^*$ are respectively the scaled and shifted versions of the student prediction $\mathbf{D}$ and the proxy label $\mathbf{D}^*$, and (s, t) are the scaling factors obtained using the least-square approach:

$$(s, t) = \arg\min_{s,t} \sum_{(x,y)\in\mathbf{M}} (s\mathbf{D} + t - \mathbf{D}^*)^2 \qquad (2)$$

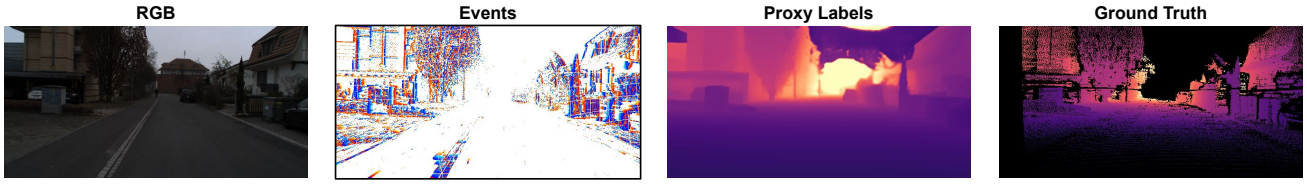

Figure 3. **Labels Distillation from Frame-Based Vision Foundation Model.** Given the availability of aligned color and event modalities, e.g., collected by a DAVIS346B sensor, we can exploit a VFM to extract proxy labels from the color images, resulting in much dense supervision compared to the one provided by semi-dense LiDAR annotations.

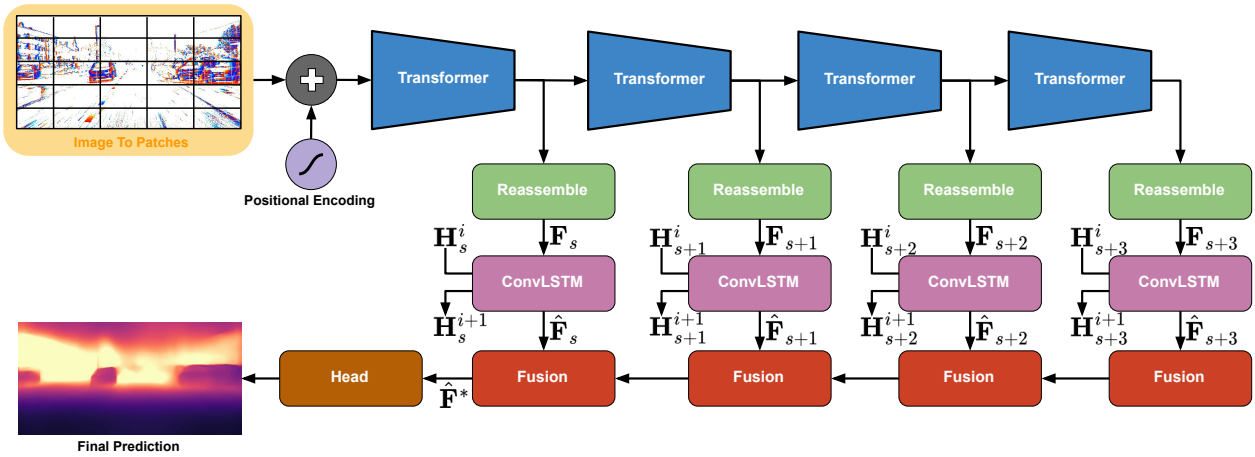

Figure 4. **Proposed Recurrent VFM.** Our DepthAnyEvent-R model processes image patches with positional encoding through multiple transformer stages that produce multi-scale feature maps $\mathbf{F}_s$. These features are combined with hidden states $\mathbf{H}_s^i$ in ConvLSTM modules $\mathcal{R}_s$ to incorporate temporal information from previous event stacks, generating enhanced feature maps $\hat{\mathbf{F}}_s$ and updated hidden states $\mathbf{H}_s^{i+1}$. A hierarchical fusion process integrates features from different scales to predict the final depth prediction $\hat{\mathbf{F}}^*$.

The regularization term $\mathcal{L}_{reg}$ is defined as follows:

$$\mathcal{L}_{reg}(\hat{\mathbf{D}}, \hat{\mathbf{D}}^*) = \sum_{k=1}^{K} \frac{1}{|\mathbf{M}_k|} \sum_{(x,y)\in\mathbf{M}_k} (|\nabla_x \mathbf{R}_k| + |\nabla_y \mathbf{R}_k|) \tag{3}$$

where $\mathbf{R}_k = \hat{\mathbf{D}}_k - \hat{\mathbf{D}}_k^*$ is the difference of maps at scale k and $\mathbf{M}_k$ is the set of valid pixels at scale k.

To ensure alignment, frame and event cameras must be calibrated – intrinsically done in the DAVIS camera – and events are sliced from the frame's acquiring timestamp.

4.2. Casting VFMs to the Event Domain

Frame-based monocular depth models cannot be used directly on events, given the diverse nature of the latter. Hence, to adapt their capabilities to the event domain, we choose an appropriate event representation that can reduce the gap between frames and events encoding. Furthermore, we exploit the sequential nature of temporal events, proposing a novel recurrent architecture of DAv2.

Choosing the Right Event Representation. The events stream contains spatial and temporal information; hence, a good event representation should capture both to ensure limited loss of information. Since monocular models naturally process RGB frames – *i.e.*, they produce a depth map given an image $\mathbf{I} \in \mathbb{R}^{W\times H\times 3}$ as input – we have to choose an event representation that encodes both spatial and temporal requirements within an RGB frame to pursue minimal modifications of the pre-trained VFM.

Purposely, the Tencode [16] representation fits with our aim. Consequently, starting from a sliced event history $\mathcal{E}_{t_d}$, either using SBT or SBN [22], Tencode encodes $\mathcal{E}_{t_d}$ into a stack $\mathbf{E}$ as follows:

$$\mathbf{E}(x_k, y_k) = \begin{cases} (1, \frac{t_d - t_k}{\Delta T}, 0) \text{ if } p_k = 1 \\ (0, \frac{t_d - t_k}{\Delta T}, 1) \text{ if } p_k = -1 \end{cases} \tag{4}$$

where $e_k = (x_k, y_k, p_k, t_k) \in \mathcal{E}_{t_d}$ is the k-th event of $\mathcal{E}_{t_d}$ and ΔT is the time interval of event slice $\mathcal{E}_{t_d}$.

VFM for Events. Although the Tencode representation significantly differs from a conventional RGB image of the same scene, we propose to adapt a pre-trained VFM to deal with the event domain through fine-tuning with event data using the Tencode representation. For this purpose, we use as the VFM a vanilla DAv2 ViT-S for our experiments. We dubbed the model as *DepthAnyEvent*.

Recurrent VFM for Events. Additionally, given the sequence nature of the event stream, Recurrent Neural Networks (RNNs) could encode previous features extracted from past event stacks into a hidden state [15, 21]. At each iteration, the recurrent module can update the hidden state

Model	Dataset	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
E2Depth [15]	MVSEC	0.527	1.122	7.894	0.512	0.244	0.363	0.637	0.811
EReFormer [21]		0.518	1.012	8.423	0.559	0.316	0.361	0.630	0.800
DepthAnyEvent		**0.466**	0.976	**7.824**	**0.480**	**0.229**	0.408	0.689	**0.847**
DepthAnyEvent-R		0.469	**0.946**	8.064	0.508	0.272	**0.428**	**0.690**	0.832
E2Depth [15]	DSEC	0.395	0.334	13.258	0.412	0.167	0.409	0.719	0.891
EReFormer [21]		0.297	0.195	11.608	0.334	0.113	0.524	0.824	0.945
DepthAnyEvent		0.297	0.186	11.072	0.330	0.108	0.519	0.827	0.948
DepthAnyEvent-R		**0.276**	**0.165**	**10.942**	**0.314**	**0.101**	**0.555**	**0.843**	**0.954**

Table 1. **Quantitative Results – Zero-Shot Generalization on MVSEC and DSEC.** All networks are trained on the EventScape synthetic dataset only, and tested without any fine-tuning.

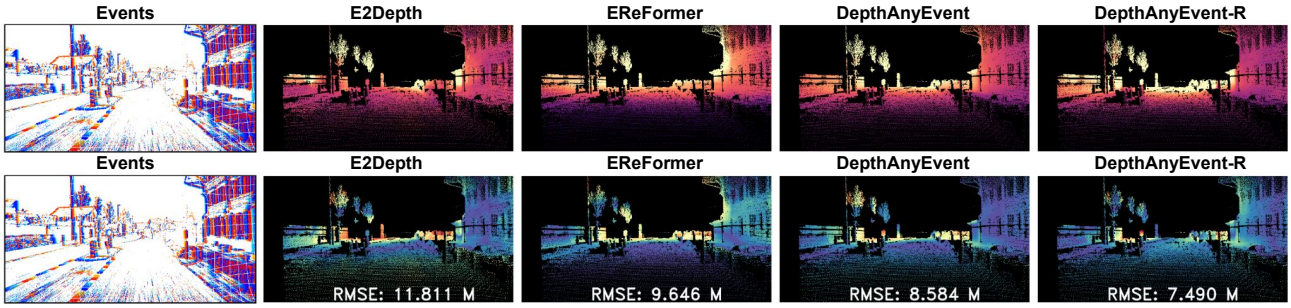

Figure 5. **Qualitative Results on DSEC dataset – Zero-Shot Generalization.** From left to right: event image, predictions by E2Depth, EReFormer, DepthAnyEvent and DepthAnyEvent-R, trained on EventScape only.

with the features extracted from the current stack, generating a new hidden state for the next iteration.

However, monocular depth models typically lack a recurrent module since they are designed to work with single-frame instances. Hence, for our purposes, this could hinder the quality of predictions, especially during static scenes where events are not triggered. To effectively adapt them to the event domain, we introduce a recurrent extension of DAv2 ViT-Small, dubbed as *DepthAnyEvent-R*, that integrates cues from previous event stacks, as outlined in Figure 4. The DAv2 architecture is composed of two main modules: a DINOv2 [24] Encoder $\mathcal{G}$ based on Visual Transformer (ViT), and a Dense Depth Decoder $\mathcal{D}$. Given an image $\mathbf{I}$ encoded with the Tencode representation, the encoder $\mathcal{G}$ first splits the image into patches and adds positional encoding to them. Next, patches are passed through multiple transformer stages and then reassembled from different stages into multi-scale feature maps $\mathbf{F}_s \in \mathbb{R}^{\frac{W}{s} \times \frac{H}{s} \times C_s}$. For each scale s, we feed the feature maps $\mathbf{F}_s$ and the hidden state $\mathbf{H}_s^i \in \mathbb{R}^{\frac{W}{s} \times \frac{H}{s} \times C_s}$ with $\mathbf{H}_s^0 = \mathbf{0}$ to a ConvLSTM [31] module $\mathcal{R}_s$ obtaining a new hidden state $\mathbf{H}_s^{i+1}$ and temporally enhanced feature maps $\hat{\mathbf{F}}_s$. Starting from the lowest scale, a series of fusion modules sequentially upsample and fuse the feature maps to obtain the final feature map $\hat{\mathbf{F}}^*$ fed to the decoder $\mathcal{D}$ to obtain the final predicted depth map.

5. Experiments

We describe our implementation details, datasets, and evaluation protocols, followed by experiments.

5.1. Implementation and Experimental Settings

Hyperparameters Settings. We set the slicing window ΔT, the number of Voxel Grid bins B, and the loss factor λ respectively to 50ms, 5, and 0.25. We implement event-based student networks E2Depth [15] and EReFormer [21] starting from their codebase. For DepthAnyEvent and DepthAnyEvent-R, we start from the DAv2 ViT-Small codebase [34]. We use PyTorch, and a single A100 GPU with 64GB of RAM. Following the original papers, we fix the learning rate to 10^{-4} and $3.2 \cdot 10^{-5}$ respectively for E2Depth and EReFormer, while we set a learning rate of $5 \cdot 10^{-6}$ for all DepthAnyEvent variants. We adjust the training steps to 75k, using the AdamW optimizer with the OneCycle scheduler, and apply data augmentations including horizontal flips and random crops at 224×224. We set the batch size to 10, except for EReFormer: given the higher memory requirements, we change it to 2. We unroll all recurrent networks – *i.e.*, E2Depth, EReFormer, and DepthAnyEvent-R – for 20 steps. We choose as the event representation Tencode [16] for DepthAnyEvent and DepthAnyEvent-R, while we maintained the original representation for E2Depth and EReFormer – *i.e.*, respectively, Voxel Grid [40] and Image-like [21]. Finally, we use the scale-invariant $\mathcal{L}$ for all networks. The settings reported are used for all experiments unless otherwise specified.

Proxy Labels Factory. We generate proxy labels from frames using the DAv2 ViT-Large trained for metric depth estimation: starting from the *Large* vanilla weights provided by the authors, we perform a fine-tuning on EventScape [8] for 10k steps with a learning rate of 10^{-6}.

Synthetic Training Setup. We obtain the synthetic

Model	Dataset	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
E2Depth [15]	MVSEC	0.420	0.806	7.268	0.455	**0.213**	0.432	0.717	0.868
EReFormer [21]		0.511	1.057	8.373	0.523	0.274	0.391	0.652	0.810
DepthAnyEvent		0.373	0.715	6.627	**0.449**	0.222	0.471	0.747	**0.884**
DepthAnyEvent-R		**0.365**	**0.691**	**6.465**	0.483	0.258	**0.489**	**0.751**	0.878
E2Depth [15]	DSEC	0.253	0.130	10.119	0.315	0.107	0.574	0.861	0.956
EReFormer [21]		0.286	0.208	11.369	0.325	0.109	0.569	0.839	0.944
DepthAnyEvent		0.201	0.079	8.880	0.266	0.077	0.664	0.917	0.975
DepthAnyEvent-R		**0.191**	**0.070**	**8.618**	**0.244**	**0.064**	**0.691**	**0.930**	**0.981**

Table 2. **Quantitative Results – In-Domain Evaluation on MVSEC and DSEC.** All networks are trained on the EventScape synthetic dataset and then further fine-tuned on MVSEC and DSEC datasets separately.

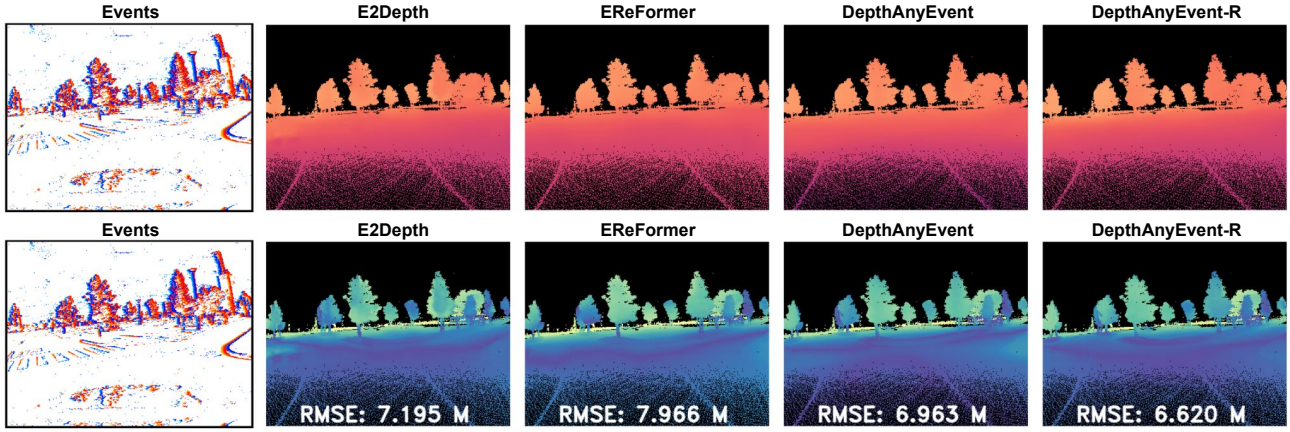

Figure 6. **Qualitative Results on MVSEC – Fine-tuned Models.** From left to right: event image, predictions by E2Depth, EReFormer, DepthAnyEvent and DepthAnyEvent-R, trained on EventScape and fine-tuned on MVSEC.

checkpoints for all networks training on the synthetic EventScape [8] dataset. While E2Depth was trained from scratch, we followed EReFormer's original paper and set Swin-T pre-trained on ImageNet as the backbone. For DepthAnyEvent and DepthAnyEvent-R, we started from the *Small* weights provided by the authors.

Fine-tuning Setup. We follow [15], fine-tuning the models to the target domain using both real and synthetic data – *i.e.*, MVSEC [39] + EventScape [8], and DSEC[9] + EventScape [8] – starting from the synthetic checkpoints obtained in the previous point.

Distillation Training Setup. We use the proxy labels previously generated with DAv2 ViT-L instead of the original sparse ground-truth. Differently from the previous point, we trained the models on the dense proxy labels only instead of a synthetic+proxy mixture.

5.2. Evaluation Datasets & Protocol

Datasets. We utilize EventScape [8] as the synthetic training set, comprising about 120k groundtruth depth maps at resolution of 512×256, captured from CARLA [4] simulator. For evaluation and domain fine-tunings we used two main benchmarks: MVSEC [39] and DSEC [9]. The dataset provides events at a resolution of 346×260 pixels from a stereo event camera consisting of two DAVIS346B sensors, which also capture spatially aligned images. ground-truth is obtained by processing data from a 16-line LiDAR using Lidar Odometry and Mapping (LOAM), yielding a total of 10k training samples and 20k testing samples. The test set is divided into a 5k-sample daytime subset and three nighttime subsets, each containing 5k samples. DSEC [9] employs two 640×480 Prophesee Gen3.1 event cameras in a stereo configuration. Ground-truth disparity is obtained using a 32-line LiDAR, processed with a Lidar Inertial Odometry algorithm, and further filtered to remove outliers. We convert the disparity ground-truth to depth based on the stereo setup parameters. Unlike MVSEC, RGB frames are captured using a pair of FLIR Blackfly S cameras. To align frames and events, we warp the RGB frames using the calibration parameters. We also apply a 640×320 center crop to mitigate misalignment artifacts in nearby objects. The dataset counts 26k training samples, divided as in [1] into 19k for training and 7k for testing.

Evaluation Metrics. We evaluate the networks using different metrics: absolute relative error (Abs Rel), square Abs Rel (Sq Rel), root mean squared error (RMSE), logarithmic RMSE (RMSE log), logarithmic scale invariant error (SI log), and accuracy with different thresholds ($\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$). We apply scale and shift to align predictions with the ground-truth before computing the metrics. We highlight using **bold** and underline the best and second best scores.

Model	Dataset	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
E2Depth Synth	MVSEC	0.527	1.122	7.894	0.512	0.244	0.363	0.637	0.811
E2Depth Distilled		**0.400**	0.817	**6.786**	0.538	0.304	**0.479**	**0.740**	0.865
E2Depth Supervised		0.420	**0.806**	7.268	**0.455**	**0.213**	0.432	0.717	**0.868**
EReFormer Synth	MVSEC	0.518	1.012	8.423	0.559	0.316	0.361	0.630	0.800
EReFormer Distilled		**0.448**	**0.817**	**7.867**	**0.498**	**0.253**	**0.434**	**0.700**	**0.842**
EReFormer Supervised		0.511	1.057	8.373	0.523	0.274	0.391	0.652	0.810
DepthAnyEvent Synth	MVSEC	0.466	0.976	7.824	0.480	0.229	0.408	0.689	0.847
DepthAnyEvent Distilled		0.397	0.771	6.910	0.495	0.260	0.461	0.735	0.870
DepthAnyEvent Supervised		**0.373**	**0.715**	**6.627**	**0.449**	**0.222**	**0.471**	**0.747**	**0.884**
DepthAnyEvent-R Synth	MVSEC	0.469	0.946	8.064	0.508	0.272	0.428	0.690	0.832
DepthAnyEvent-R Distilled		0.399	0.781	6.830	0.509	0.281	0.462	0.735	0.866
DepthAnyEvent-R Supervised		**0.365**	**0.691**	**6.465**	**0.483**	**0.258**	**0.489**	**0.751**	**0.878**
E2Depth Synth	DSEC	0.395	0.334	13.258	0.412	0.167	0.409	0.719	0.891
E2Depth Distilled		0.272	0.153	10.579	**0.309**	**0.096**	0.551	0.851	**0.959**
E2Depth Supervised		**0.253**	**0.130**	**10.119**	0.315	0.107	**0.574**	**0.861**	0.956
EReFormer Synth	DSEC	0.297	**0.195**	11.608	0.334	0.113	0.524	0.824	**0.945**
EReFormer Distilled		**0.285**	0.198	11.407	0.327	0.111	0.563	**0.839**	0.944
EReFormer Supervised		0.286	0.208	**11.369**	**0.325**	**0.109**	**0.569**	0.839	0.944
DepthAnyEvent Synth	DSEC	0.297	0.186	11.072	0.330	0.108	0.519	0.827	0.948
DepthAnyEvent Distilled		0.213	0.095	8.930	**0.253**	**0.065**	0.662	0.915	**0.980**
DepthAnyEvent Supervised		**0.201**	**0.079**	**8.880**	0.266	0.077	**0.664**	**0.917**	0.975
DepthAnyEvent-R Synth	DSEC	0.276	0.165	10.942	0.314	0.101	0.555	0.843	0.954
DepthAnyEvent-R Distilled		0.226	0.111	9.310	0.266	0.072	0.638	0.906	0.977
DepthAnyEvent-R Supervised		**0.191**	**0.070**	**8.618**	**0.244**	**0.064**	**0.691**	**0.930**	**0.981**

Table 3. **Quantitative Results – Supervised vs Distilled Models on MVSEC and DSEC.** All networks are trained on the EventScape synthetic dataset and then fine-tuned on MVSEC and DSEC datasets separately, either through distillation or on ground-truth depth labels.

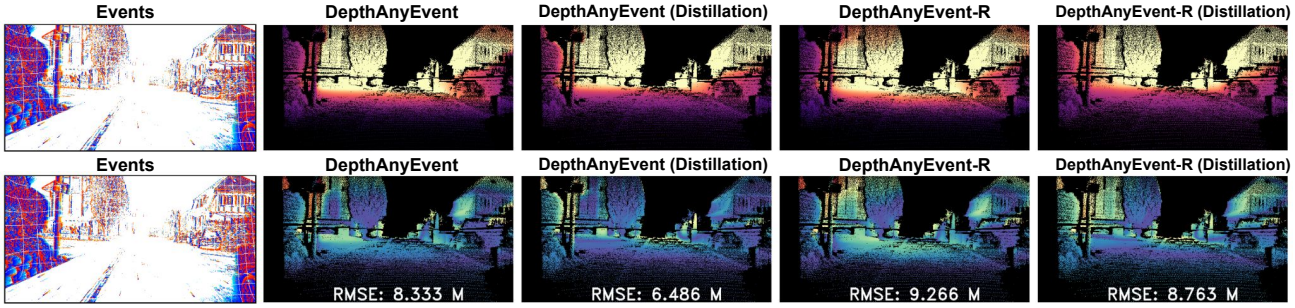

Figure 7. **Qualitative Results on DSEC – Supervised vs Distilled Models.** From left to right: event image, predictions by DepthAnyEvent and its distilled counterpart, and by DepthAnyEvent-R and its distilled counterpart.

5.3. Synthetic-to-Real Generalization

We start by evaluating the capability of the different depth estimation models to generalize from synthetic data to real event streams. Purposely, we train E2Depth, EReFormer, DepthAnyEvent, and DepthAnyEvent-R on EventScape and measure their accuracy on both MVSEC and DSEC datasets. Table 1 collects the outcome of this experiment. DepthAnyEvent and DepthAnyEvent-R achieve the best results on almost any metric, hinting how the web-scale training infused in the weights we used to initialize these models represents a solid prior for depth estimation, although coming from images, i.e., a completely different modality with respect to event streams. The two models achieve mixed results one against the other on MVSEC, while DepthAnyEvent-R consistently achieves the best generalization results over DSEC, giving a first intuition about the effectiveness of our design choice to deal with streamed event data. Figure 5 presents a qualitative comparison of predictions from different models, showcasing the superior zero-shot capabilities of our DepthAnyEvent and DepthAnyEvent-R models.

5.4. Supervised Fine-tuning

We now evaluate the accuracy of each model when trained on real event data annotated with semi-dense ground-truth depth. To this aim, we take the weights obtained after training on EventScape and perform further fine-tuning on MVSEC and DSEC separately, then evaluating on the corresponding validation sets. Table 2 reports the results of this evaluation. We can notice, once again, the notable gap in performance between DepthAnyEvent and DepthAnyEvent-R against existing methods EReFormer and E2Depth, confirming again the strong advantage that our models can exploit from the cross-modal training being conducted for image-based depth estimation. Specifically, this time we can notice how DepthAnyEvent-R consistently outperforms the vanilla DepthAnyEvent model on both MVSEC and DSEC datasets, validating our proposed design tailored to event-based depth estimation.

Figure 6 shows a qualitative comparison between the

Model	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
E2Depth [15]	0.344	0.253	13.467	0.376	0.098	0.447	0.755	0.915
EReFormer [21]	0.387	0.401	13.954	0.395	0.124	0.486	0.776	0.892
DepthAnyEvent	0.277	0.170	11.117	0.292	0.051	0.585	0.860	0.955
DepthAnyEvent-R	**0.252**	**0.128**	**9.824**	**0.268**	**0.045**	**0.592**	**0.900**	**0.971**

Table 4. **Metric Depth Evaluation.** Training and evaluation on DSEC dataset.

	Model	Supervision	Experiment	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
(A)	DepthAnyEvent-R	Distillation	Tencode+DAv2	**0.399**	**0.781**	**6.830**	0.509	0.281	**0.462**	**0.735**	0.866
(B)		Distillation	Tencode+DepthPro	0.429	0.942	7.472	**0.452**	**0.208**	0.444	0.726	**0.869**
(C)	DepthAnyEvent-R	Ground-truth	Tencode+DAv2	0.365	**0.691**	**6.465**	0.483	0.258	0.489	0.751	0.878
(D)		Ground-truth	VoxelGrid+DAv2	0.382	0.719	6.932	0.444	0.215	0.473	0.742	0.877
(E)		Ground-truth	Tencode+DAv2 (no pretrain)	0.446	0.799	7.492	0.506	0.260	0.390	0.678	0.845
(F)		Ground-truth + Distillation	Tencode+DAv2	**0.362**	0.697	6.511	**0.438**	**0.211**	**0.494**	**0.760**	**0.890**

Table 5. **Ablation Studies.** Training and evaluation on MVSEC dataset.

Model	Inference (ms)	Memory (MB)
E2Depth [15]	1.50	242
EReFormer [21]	35.75	534
DepthAnyEvent	**1.26**	**71**
DepthAnyEvent-R	9.20	202

Table 6. **Computational Analysis.** Inference time on A100 GPU.

predictions by the different models, highlighting the superior accuracy achieved by DepthAnyEvent and, even higher, by DepthAnyEvent-R.

5.5. Cross-Modal Distillation

We now assess the effectiveness of our cross-modal distillation strategy compared to conventional, supervised training requiring the availability of costly depth annotations from active sensors. Table 3 collects the results achieved by each model under the training configuration considered so far, as well as after being trained according to our distillation approach. In most cases, we can notice how the models trained through distillation are comparable, and sometimes even better than their supervised counterparts.

Figure 7 show some qualitative examples from the DSEC dataset, comparing the predictions by DepthAnyEvent and DepthAnyEvent-R when trained with ground-truth or through distillation. In both cases, distilled models are even more accurate than those supervised with ground-truth.

5.6. Metric Depth Evaluation

Finally, we assess the accuracy of our models when trained to predict metric rather than affine-invariant depth. Table 4 collects the results achieved by existing networks and ours when trained on the DSEC dataset for metric depth prediction, evaluated on the validation set of the very same dataset. We can appreciate how our two architectures achieve the best results, with DepthAnyEvent-R consistently yielding the best results on any evaluation metrics.

5.7. Ablation Studies

We conclude with a study about the impact of different modules in our framework. In the former case, we train different instances of DepthAnyEvent-R on the MVSEC dataset and evaluate on its validation set. Results are collected in Table 5, with row (A) representing the configuration used in the previous experiments.

Different VFMs for distillation. Row (B) shows that replacing Depth Anything v2 with a different VFM for distillation – i.e., Depth Pro – yields close results, although slightly worse on most metrics.

Input representation. In rows (C) and (D), we report the results achieved by training our model with ground-truth labels, when processing either Tencode or a voxel-grid representation used to encode raw events. The former yields almost consistently better results.

Pre-training. By training our model starting from DAv2 pretrained weights, we can greatly improve its performance. Indeed, when training DepthAnyEvent-R from scratch (E), the accuracy consistently drops.

Combining distillation with ground-truth labels. Finally, we show how deploying both our cross-modal distillation paradigm and ground-truth annotations (when available) further improves the final model on most metrics.

5.8. Runtime and Memory Requirements

Table 6 reports a computational analysis for any model involved in our evaluation. DepthAnyEvent achieves the fastest predictions, using as few as 80MB for a single inference. E2Depth exposes a very similar inference time, although requiring nearly 4× the memory, while EReFormer runs consistently slower and increases the memory usage to up to 0.5GB. Compared to DepthAnyEvent, the DepthAnyEvent-R variant runs slower, yet still in real-time, and yields more accurate predictions.

6. Conclusions

In this paper, we presented a novel approach to event-based monocular depth estimation that leverages the power of pre-trained Visual Foundation Models. Our cross-modal distillation strategy effectively transfers knowledge from frame-based models to the event domain, addressing the crucial challenge of limited ground truth data for event cameras. Experimental results with synthetic and real-world

datasets validate our method, showing competitive performance compared to fully supervised methods without requiring expensive depth annotations. Moreover, we have demonstrated two effective methods for adapting VFMs to event data: a vanilla adaptation and a recurrent architecture that better captures the nature of event streams, yielding state-of-the-art performance.

Acknowledgment. This study was carried out within the MOST – Sustainable Mobility National Research Center and received funding from the European Union Next-GenerationEU – PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1033 17/06/2022, CN00000023. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

We also acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support.

References

[1] Luca Bartolomei, Matteo Poggi, Andrea Conti, and Stefano Mattoccia. Lidar-event stereo fusion with hallucinations. In *European Conference on Computer Vision*, pages 125–145. Springer, 2024. 3, 6

[2] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4016–4023. IEEE, 2023. 3

[3] Mingyue Cui, Yuzhang Zhu, Yechang Liu, Yunchao Liu, Gang Chen, and Kai Huang. Dense depth-map estimation based on fusion of event camera and sparse lidar. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022. 2

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 6

[5] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 2

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2

[7] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 154–180, 2022. 2

[8] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. 2, 3, 5, 6

[9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954, 2021. 3, 6

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[11] Suman Ghosh and Guillermo Gallego. Event-based stereo depth estimation: A survey. *arXiv preprint arXiv:2409.17680*, 2024. 3

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016. 2

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018. 2

[14] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 2

[15] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. *CoRR*, abs/2010.08350, 2020. 1, 2, 3, 4, 5, 6

[16] Ze Huang, Li Sun, Cheng Zhao, Song Li, and Songzhi Su. Eventpoint: Self-supervised interest point detection and description for event-based camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5396–5405, 2023. 3, 4, 5

[17] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. In *ICCV*, 2023. 2

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[19] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *CoRR*, abs/1907.01341, 2019. 3

[20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

[21] Xu Liu, Jianing Li, Jinqiao Shi, Xiaopeng Fan, Yonghong Tian, and Debin Zhao. Event-based monocular depth estimation with recurrent transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7417–7429, 2024. 2, 3, 4, 5, 6

[22] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6114–6123, 2022. 4

[23] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 5

[25] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 2

[26] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 2

[27] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009. 2

[28] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 2

[29] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. CED: color event camera dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2019. 2

[30] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. 2

[31] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 5

[32] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 2

[33] 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*, 2024. 2

[34] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 2, 5

[35] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 2

[36] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. 2

[37] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pages 668–678. IEEE, 2022. 2

[38] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619, 2017. 2

[39] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3 (3):2032–2039, 2018. 3, 6

[40] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. *CoRR*, abs/1812.08156, 2018. 2, 3, 5

[41] Junyu Zhu, Lina Liu, Bofeng Jiang, Feng Wen, Hongbo Zhang, Wanlong Li, and Yong Liu. Self-supervised event-based monocular depth estimation using cross-modal consistency, 2024. 3