

Depth Any Event Stream: Enhancing Event-based Monocular Depth Estimation via Dense-to-Sparse Distillation

Jinjing Zhu^{1*} Tianbo Pan^{1*} Zidong Cao¹ Yexin Liu² James T. Kwok² Hui Xiong^{1†}

¹ HKUST(GZ) ² HKUST

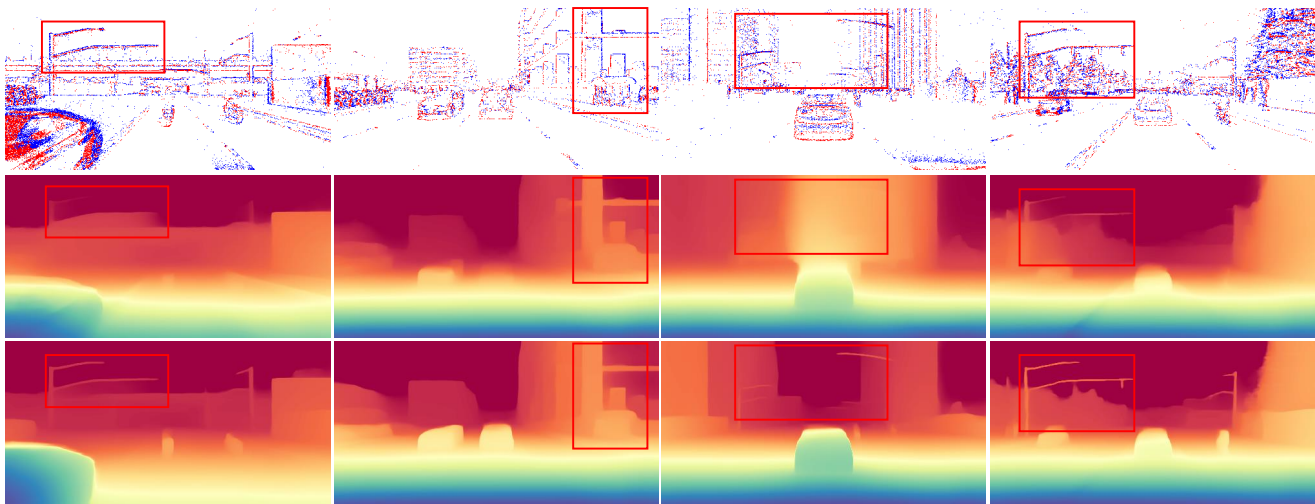


Figure 1. Our event-based monocular depth estimation model processes only sparse event data, yielding highly accurate depth estimations. **Top:** Input event data. **Middle:** DAM [41] with input event data. **Bottom:** EventDAM with input event data.

Abstract

With the superior sensitivity of event cameras to high-speed motion and extreme lighting conditions, event-based monocular depth estimation has gained popularity to predict structural information about surrounding scenes in challenging environments. However, the scarcity of labeled event data constrains prior supervised learning methods. Unleashing the promising potential of the existing RGB-based depth foundation model, DAM [41], we propose Depth Any Event stream (**EventDAM**) to achieve high-performance event-based monocular depth estimation in an annotation-free manner. EventDAM effectively combines paired dense RGB images with sparse event data by incorporating three key cross-modality components: Sparsity-aware Feature Mixture (**SFM**), Sparsity-aware Feature Distillation (**SFD**), and Sparsity-invariant Consistency Module (**SCM**). With the proposed sparsity metric, SFM mixes features from RGB images

and event data to generate auxiliary depth predictions, while SFD facilitates adaptive feature distillation. Furthermore, SCM ensures output consistency across varying sparsity levels in event data, thereby endowing EventDAM with zero-shot capabilities across diverse scenes. Extensive experiments across a variety of benchmark datasets, compared to approaches using diverse input modalities, robustly substantiate the generalization and zero-shot capabilities of EventDAM.

1. Introduction

Compared with traditional RGB image sensors, event cameras [7] offer several distinct advantages, including high temporal resolution, high dynamic range, and markedly reduced latency. It excels at capturing precise structural information in challenging conditions, such as poorly illuminated areas [27] and environments with moving objects at high speed [39]. Recently, increasing research has shown considerable potential in areas such as classification [43],

*These authors contributed equally to this work.

†Corresponding author.

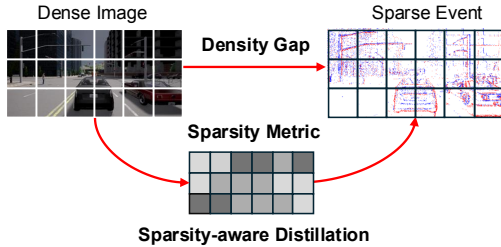


Figure 2. EventDAM aims to distill DAM into an event-based depth estimation model, taking into account the sparsity of event data.

object detection [25, 31], tracking [4, 8, 42], and segmentation [5, 20, 21, 23]. In particular, event-based depth estimation [6, 11, 16, 30, 36] holds considerable appeal due to its substantial benefits for applications such as robotic navigation and autonomous driving [16, 33, 38]. While these methods [14, 30] successfully take full advantage of RGB images and event data to complement each other, there remains a major challenge: *the limited size and diversity of annotated datasets*. This bottleneck restricts the ability of event-based depth estimation methods to generalize effectively to the complexity and variability of real-world scenes, thereby limiting their broader applicability.

Recent advances in vision foundation models, such as Segment Anything Model (SAM) [22], DINO [28], and Depth Anything Model (DAM) [41], are revolutionizing the field of computer vision, driven by the availability of large-scale labeled datasets. This progress holds substantial promise for exploring the applicability of vision foundation models to the event domain [5, 23]. For instance, EventSAM [5] adapts SAM for universal object segmentation with event data, while OpenESS [23] combines CLIP [32] and SAM to enable open-vocabulary event-based semantic segmentation. Inspired by these developments, we aim to leverage the extensive knowledge embedded in large-scale RGB image datasets and the advanced learning capabilities of off-the-shelf models, such as DAM, to drive progress in event-based depth estimation. Notably, an increasing number of datasets containing paired RGB image and event data provide a unique opportunity for this endeavor [10, 12]. Therefore, we propose adapting DAM to the event domain, unlocking the potential of DAM for event-based depth estimation.

Nevertheless, due to the distinctive sensing mechanisms of RGB and event cameras, inherent differences between RGB images and event data bring the severe challenges. Specifically, as shown in Fig. 2, a significant density discrepancy between the dense RGB image and sparse event data limits the direct application of DAM in the event domain (See Fig. 1). To tackle this, we propose Depth Any Event stream (**EventDAM**) to achieve high-performance event-based monocular depth estimation in an annotation-free manner. With explicit consideration of the sparsity of event data, we propose three core components to facili-

tate effective knowledge transfer from DAM to event-based DAM. First, we introduce Sparsity-aware Feature Mixture (**SFM**), which bridges the density gap by mixing features from both RGB images and event data using the sparsity metric, generating auxiliary depth predictions to guide the student model’s training (*cf.* Sec. 3.3). The sparsity metric quantifies patch-level registered information across regions in the event data, assigning lower sparsity values to regions with richer informativeness and higher values to sparser regions with fewer triggered events (See Fig. 2). Second, to further facilitate the dense-to-sparse distillation, we develop Sparsity-aware Feature Distillation (**SFD**) with the sparsity metric. This approach leverages sparsity-based weighting to guide the distillation from teacher features into student features, considering the triggered events of relevant regions, rather than directly distilling features (*cf.* Sec.3.4). Finally, to enhance the model’s zero-shot capability, we present a Sparsity-invariant Consistency Module (**SCM**). This module maintains output consistency across varying sparsity levels in event data, improving the model’s generalization across diverse scenes (*cf.* Sec. 3.5).

We conduct extensive experiments to evaluate the effectiveness of our EventDAM on two benchmark datasets, EventScape (See Fig. 1) and MVSEC [45]. Furthermore, we assess EventDAM’s zero-shot capability by training it on the EventScape dataset and testing its performance on two other datasets: DENSE [16] and MVSEC. The qualitative and quantitative results demonstrate that EventDAM significantly outperforms current state-of-the-art (SOTA) event and image-based methods on both EventScape and DENSE.

In summary, our contributions are as follows: **(I)** We introduce EventDAM, *a pioneering approach* that distills DAM via dense-to-sparse distillation for enhancing event-based depth estimation. **(II)** We propose novel techniques, including Sparsity-aware Feature Mixture (SFM), Sparsity-aware Feature Distillation (SFD), and a sparsity-invariant Consistency Module (SCM), to enable effective knowledge transfer from DAM to event-based DAM. **(III)** Comprehensive experiments validate the effectiveness of the proposed EventDAM, highlighting its superior performance and its zero-shot capability across diverse scenes.

2. Related Work

Event-based monocular depth estimation. Event-based monocular depth estimation seeks to generate pixel-wise scene depth by using event data [37, 44]. The pioneering work, E2Depth [16], adopts recurrent convolutional neural networks to produce dense depth maps from asynchronous event streams. Subsequently, approaches integrating RGB and event data have been proposed to improve depth estimation accuracy [6, 11, 30, 36]. For instance, ER-F2D [6] employs a unified transformer to capture inter-modal dependencies, thereby enhancing accuracy, while SRFNet [30]

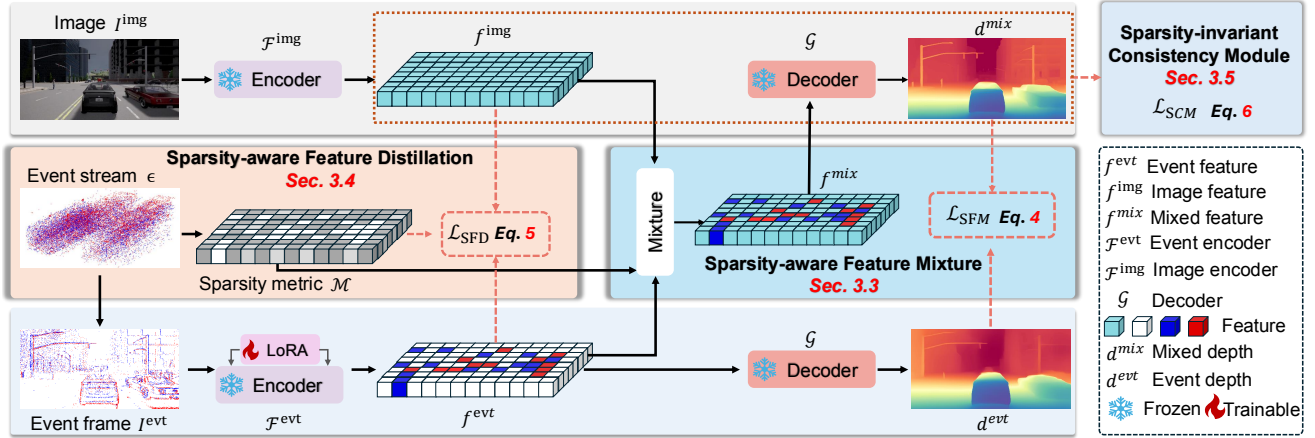


Figure 3. **Overview of the EventDAM framework.** We distill off-the-shelf knowledge from DAM for event data. Given paired event data I^{evt} and RGB images I^{img} , we extract their features from the event encoder \mathcal{F}^{evt} and image encoder \mathcal{F}^{img} . To enhance the distillation from DAM into event-based DAM, we propose **Sparsity-aware Feature Mixture (SFM)** to mix their features to generate the auxiliary depth d^{mix} with depth decoder \mathcal{G} (cf. Sec. 3.3), **Sparsity-aware Feature Distillation (SFD)** to enable adaptive feature distillation with the sparsity metric (cf. Sec. 3.4), and **Sparsity-invariant Consistency Module (SCM)**, which generates sparser event data (cf. Sec. 3.5).

leverages spatial sparsity to harness the strengths of both modalities. *Despite these advancements, a significant limitation in prior work is the reliance on annotated datasets, which are costly and time-intensive to produce. To address it, we propose utilizing DAM trained on large-scale image datasets to improve performance with event data, aiming to develop a high-performance event-based depth estimation model without requiring annotated event data.*

Cross-modal knowledge distillation. Knowledge distillation (KD) is proposed for transferring knowledge from one neural network to another [17, 40, 46, 47]. Building upon this foundation, cross-modal KD [19, 24] has emerged to facilitate knowledge transfer across different modalities, aiming to leverage insights derived from a large-scale labeled dataset in one modality to benefit another. For example, Gupta et al. [13] employs mid-level representations from a labeled modality to supervise the learning of representations in a paired, unlabeled modality. More recent studies [1, 48] address the complex challenge of transferring knowledge from a source modality to a target modality without direct access to task-relevant data from the source. *In this study, we strive to utilize RGB-event pairs to harness the capabilities of DAM to enhance event-based depth estimation, specifically by bridging the density gap between dense RGB images and sparse event data.*

Vision foundation models (VFMs). VFMs fundamentally transform the field of artificial intelligence by leveraging vast amounts of data [22, 32, 41] and self-supervised learning [3, 15, 29]. CLIP [32], trained on a large-scale dataset of image-text pairs, demonstrates strong zero-shot transfer capabilities across various downstream tasks. SAM [22], leveraging a dataset of 11 million diverse images, exhibits robust zero-shot instance segmentation performance. DINO [29] investigates the potential of self-supervised learning to learn

general-purpose visual features from a substantial amount of curated data. Building on these advancements in large VFMs, recent research has explored adapting foundation models for alternative modalities to fully exploit their capabilities. EventSAM [5] extends SAM for universal object segmentation in event-based data, while OpenESS [23] integrates CLIP and SAM to enable open-vocabulary, event-based semantic segmentation. *Differently, our work investigates the potential of DAM [41] to improve event-based depth estimation through distillation from DAM to event-based DAM.*

3. Method

3.1. Overview

Our study serves as an initial effort to leverage DAM [41] to enhance event-based depth estimation *without requiring access to ground-truth depth data*. We begin with a comprehensive description of our proposed EventDAM framework, outlining its components, including the inputs, event frame-like representations, feature encoding, and fine-tuning strategy (cf. Sec. 3.2). To facilitate effective distillation from DAM to event-based DAM, we introduce three key components: Sparsity-aware Feature Mixture (cf. Sec. 3.3), Sparsity-aware Feature Distillation (cf. Sec. 3.4), and Sparsity-invariant Consistency Module (cf. Sec. 3.5). Fig. 3 provides an overview of EventDAM, with each component described in detail in the following sections.

3.2. Event-based Monocular Depth Estimation

Inputs. Given a set of events captured by an event camera, our objective is to estimate the depth of event data, denoted as $e = (x, y, t, p)$, where (x, y) represents the pixel location, t denotes the timestamp of the observed change, and $p \in \{+1, -1\}$ indicates the polarity of the event, corresponding

to an increase or decrease in brightness, respectively. An event is triggered when the change in logarithmic brightness surpasses a predefined threshold, allowing the event camera to capture events asynchronously. Simultaneously, an RGB camera acquires conventional images, $I^{\text{img}} \in \mathbb{R}^{3 \times H \times W}$, where H and W denote the spatial resolution.

Event frame-like representation. Due to the sparsity, high temporal resolution, and asynchronous nature of event streams, it is common to convert raw event stream ε into more regular representations $I^{\text{evt}} \in \mathbb{R}^{C \times H \times W}$. Therefore, in this work, we convert the event stream into a sequence of frames. Specifically, we adopt the approach outlined in prior studies [9, 10, 16, 30, 35], where we divide the event stream into events ϵ that occur within a fixed time interval ΔT . We then transform each event stream into a frame-like representation $F \in \mathbb{R}^{B \times 2 \times H \times W}$. Each location (x, y) in F is represented with two histogram-like vectors of B bins. Each histogram discretizes ΔT in each bin and counts the number of positive or negative events occurring in the corresponding period $\Delta T/B$. In this work, we choose $\Delta T = 50\text{ms}$ and $B = 5$ temporal bins. To make the frame-like representation denser, we aggregate the counts in each pixel of B bins and then transform the final representations $\hat{F} \in \mathbb{R}^{2 \times H \times W}$ into event frame $I^{\text{evt}} \in \mathbb{R}^{3 \times H \times W}$ with spatial size $H \times W$.

Feature encoding. Each image or event frame is partitioned into L non-overlapping $P \times P$ square patches, where $L = H \times W/P^2$. Let $\mathcal{F}^{\text{evt}} : \mathbb{R}^{3 \times H \times W} \rightarrow \mathbb{R}^{D \times \hat{H} \times \hat{W}}$ denote a student encoder that processes an event embedding I^{evt} and outputs a D -dimensional event feature f^{evt} with spatial dimensions \hat{H} and \hat{W} , where $\hat{H} = H/P$ and $\hat{W} = W/P$. Likewise, the teacher encoder \mathcal{F}^{img} outputs a D -dimensional image feature f^{img} with the same downsampled spatial dimensions, \hat{H} and \hat{W} . Furthermore, we employ a depth decoder \mathcal{G} , which inputs the event feature f^{evt} and image feature f^{img} to generate depth predictions d^{evt} and d^{img} , respectively.

Fine-tuning strategy. To preserve the strong generalization and zero-shot capability of DAM [41] while adapting it to event data, we fine-tune its encoder using Low-Rank Adaptation (LoRA) [18] and freeze the decoder, rather than training the entire model (as demonstrated in Tab. 8). With this training strategy, the event-based DAM can be implemented and trained as shown in Fig. 3.

3.3. Sparsity-aware Feature Mixture

Due to the inherent density discrepancy between dense RGB images and sparse event data, directly distilling knowledge from the DAM, acting as a teacher, to the event-based DAM, serving as a student, may be suboptimal for depth estimation. To address this challenge, we introduce the Sparsity-aware Feature Mixture (SFM) as a mechanism to derive auxiliary depth prediction that effectively bridges the gap between these two modalities. Specifically, we propose mixing image

and event features to construct an auxiliary depth prediction that can guide the training process of the event-based DAM. We utilize the representation \hat{F} to derive a sparsity metric \mathcal{M} . This metric enables the assessment of informativeness across different areas, where regions with richer informativeness are assigned lower sparsity values, and sparser regions, characterized by fewer triggered events, receive higher sparsity values. More concretely, we partition the representation \hat{F} into patches of size $P \times P$, and calculate the number of triggered events, N_i , within i -th patch. The sparsity metric \mathcal{M}_i for i -th patch is defined as

$$\mathcal{M}_i = 1 - \frac{N_i}{P \times P}. \quad (1)$$

After calculating the metric \mathcal{M}_i for the i -th patch, we proceed by selecting the i -th event feature, f_i^{evt} , extracted from the i -th patch using the event-based DAM, where the corresponding \mathcal{M}_i values below the median sparsity threshold for the given input. This selection criterion is based on the observation that higher sparsity values indicate a larger density gap between the event and image regions. Consequently, our aim is to distill knowledge from image regions to event regions with more informativeness, which are typically characterized by lower sparsity. Following this, we select the corresponding image features, f^{img} , from these unselected regions and combine them with the event features, f^{evt} , resulting in the mixed features, f^{mix} . Since the mixed features are derived from image regions and event regions with fewer triggered events, the gap between the event features f^{evt} and the mixed features f^{mix} is relatively small. These mixed features, f^{mix} , are then fed into the decoder, \mathcal{G} , to generate the auxiliary depth prediction, d^{mix} , which serves as a ground-truth depth to guide the training of the event-based DAM (Fig. 3).

As noted by DAM [41], we first normalize the depth values (d^{mix} and d^{evt}) to $0 \sim 1$ on each depth map. We then apply an affine-invariant loss to ignore the unknown scale and shift in each sample:

$$\hat{d}^{\text{evt}} = \frac{d_i^{\text{evt}} - t(d^{\text{evt}})}{s(d^{\text{evt}})}, \quad (2)$$

where $t(d^{\text{evt}})$ and $s(d^{\text{evt}})$ is applied to align the depth predictions to have zero translation and unit scale:

$$t(d^{\text{evt}}) = \text{median}(d^{\text{evt}}), s(d^{\text{evt}}) = \frac{1}{HW} \sum_{i=1}^{HW} |d_i^{\text{evt}} - t(d^{\text{evt}})|. \quad (3)$$

Similarly, the prediction d^{mix} is scaled and shifted to obtain \hat{d}^{mix} . To guide the training of the event-based DAM, we employ the scale-invariant log loss $\mathcal{L}_{\text{silog}}$, as described in [2, 26], to ensure robust and scale-consistent depth estimation. Furthermore, we apply the multi-scale scale-invariant gradient matching loss \mathcal{L}_{gsm} [34], which achieves both shift-

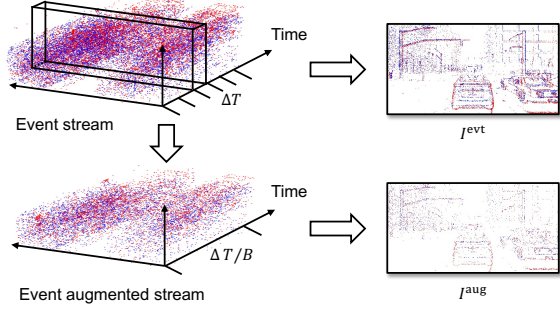


Figure 4. The augmented representation, denoted as I^{aug} , is derived by randomly extracting a single bin from the raw event stream. This representation I^{aug} is subsequently utilized to enable the distillation process, transferring knowledge from a dense image to a sparser augmented event data.

invariance and scale-invariance in disparity space. The objective of SFM is defined as:

$$\mathcal{L}_{SFM} = \mathcal{L}_{silog}(\hat{d}^{evt}, \hat{d}^{mix}) + \mathcal{L}_{gm}(\hat{d}^{evt}, \hat{d}^{mix}). \quad (4)$$

3.4. Sparsity-aware Feature Distillation

In addition to facilitating knowledge transfer through SFM in the prediction space, we introduce Sparsity-aware Feature Distillation (SFD) to enable distillation within the feature space. The key idea behind SFD is to leverage the sparsity metric, denoted as \mathcal{M} , to guide the distillation process of image features into event features. This is achieved by dynamically adjusting the weight assigned to each region based on its sparsity. The approach emphasizes transferring knowledge from image features to event features within denser regions, while mitigating the negative impact of direct transfer from image features to event features of sparse regions, which may lead to suboptimal performance. To quantify the discrepancies between features, we utilize cosine similarity, $\cos(\cdot)$, while incorporating the sparsity metric \mathcal{M} to modulate the distillation process. This method ensures that the distillation prioritizes more informative and denser regions in the event data, thereby improving the effectiveness of cross-modal knowledge transfer. And the objective of SFD is defined as:

$$\mathcal{L}_{SFD} = \sum_{i=1}^{\hat{H}\hat{W}} (1 - \mathcal{M}_i) \left(1 - \cos(f_i^{evt}, f_i^{img}) \right), \quad (5)$$

where f_i^{evt} denotes the event features and f_i^{img} denotes the image features. The sparsity metric \mathcal{M} serves as the weight for feature similarity, effectively directing the distillation process towards the denser regions of the event data.

3.5. Sparsity-invariant Consistency Module

Although SFM and SFD enable EventDAM to harness the potential of DAM for event data, its performance may degrade under real-world conditions where event data becomes sparse. Event data is inherently sparse, noisy, and irregular, which can impede the generalization of the proposed

approach across different domains or scenarios. To address this concern, we propose a Sparsity-invariant Consistency Module (SCM) (illustrated in Fig. 4), which enforces consistency of feature and depth predictions across varying levels of sparsity in the event data. The key motivation is that, regardless of the degree of sparsity in the event data, features and depth predictions for the same scene should remain consistent. The SCM is designed to enhance the generalization of EventDAM in sparse event scenes, ensuring sustained predictive accuracy across varying levels of sparsity and challenging conditions.

Specifically, we begin by randomly extracting one temporal bin with a time span of $\Delta T/B$ from the raw event stream to generate an augmented event stream, which is subsequently transformed into an event frame-like representation, I^{aug} . Based on the augmented event stream, we compute the sparsity metric \mathcal{M}^{aug} and subsequently obtain the augmented event feature f^{aug} using the event encoder \mathcal{F}^{evt} . The feature f^{aug} is then fed into decoder \mathcal{G} to produce the augmented depth prediction d^{aug} . Finally, both the augmented event feature f^{aug} and augmented depth prediction d^{aug} are processed through the SFM and SFD alongside the corresponding image I^{img} . The objective of the SCM is defined as follows:

$$\begin{aligned} \mathcal{L}_{SCM} = & \mathcal{L}_{silog}(\hat{d}^{aug}, \tilde{d}^{mix}) + \mathcal{L}_{gm}(\hat{d}^{aug}, \tilde{d}^{mix}) \\ & + \sum_{i=1}^{\hat{H}\hat{W}} (1 - \mathcal{M}_i^{aug}) \left(1 - \cos(f_i^{aug}, f_i^{img}) \right), \end{aligned} \quad (6)$$

where \hat{d}^{aug} represents the scaled and shifted version of the prediction d^{aug} , while \tilde{d}^{mix} denotes the scaled and shifted version of the mixed prediction derived from the combined feature generated by mixing f^{aug} and f^{img} .

Connecting all the pieces above, the total objective of EventDAM is formulated as:

$$\mathcal{L} = \mathcal{L}_{SFM} + \lambda_{SFD} \mathcal{L}_{SFD} + \lambda_{SCM} \mathcal{L}_{SCM}, \quad (7)$$

where λ_{SFD} and λ_{SCM} represent hyper-parameters used to balance loss terms, set to values of 2 and 0.2, respectively.

4. Experiment

4.1. Settings

Datasets. To comprehensively explore the potential of DAM [41] for event-based depth estimation, we fine-tune and evaluate EventDAM on EventScape [10] and MVSEC [45] datasets, respectively. To further explore the zero-shot capability of EventDAM, we train it on the EventScape dataset and test its performance on two other datasets: DENSE [16] and MVSEC [45].

Evaluation metrics. Following [6, 14, 30], we measure average absolute depth errors at cut-off depths of 10m, 20m, and 30m. We further evaluate our method with three percentage metrics δ_i , where $i \in \{1.25, 1.25^2, 1.25^3\}$.

Method	Input	10m ↓	20m ↓	30m ↓
E2Depth [16]	E	1.79	5.35	8.31
RAMNet [11]	E+I	0.81	2.26	3.58
HMNet [14]		0.55	1.80	3.27
ER-F2D [6]		0.67	1.69	2.81
SRFNet [30]		1.27	1.68	2.76
EventDAM-S	E	0.59	1.60	2.47
EventDAM-B		0.54	1.54	2.32
EventDAM-L		0.56	1.52	2.30

Table 1. Absolute mean depth error results on EventScape (in meters). (E) implies that event data is adopted as the input, and (E+I) means both event and image are adopted. Best results are highlighted in **Red** while runner-up in **Blue**.

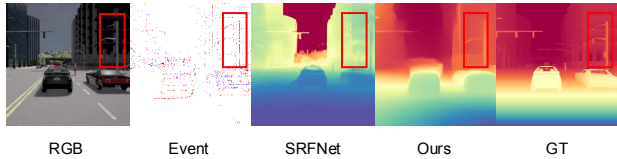


Figure 5. Qualitative comparison between SRFNet and EventDAM.

Implementation details. Following SRFNet, we center-crop the input image and event frame dimensions to 252×504 pixels to ensure consistency with ViT input settings. For the encoder, we utilize pre-trained models DAM V2 [41] of the Vision Transformer (ViT) in three configurations: small (ViT-S), base (ViT-B), and large (ViT-L). During training, we employ a batch size of 16 for ViT-S and ViT-B, while a batch size of 8 is used for ViT-L, corresponding to their respective computational demands. Additionally, we implement LoRA with a rank $r=4$ to fine-tune the encoder in our EventDAM. We train EventDAM with a learning rate of 1×10^{-4} , using the Adam optimizer for 5 epochs. All experiments are conducted on a single NVIDIA A100 80GB GPU.

Comparing methods. To thoroughly evaluate the effectiveness of the proposed EventDAM, we conduct a comparative analysis against SOTA event-based depth estimation methods. Specifically, E2Depth [16] employs only event data for depth estimation, while RAMNet [11], EvT⁺ [36], HMNet [14], ERF2D [6], and SRFNet [30] utilize **paired event and RGB** data to predict depth.

4.2. Experiment Results

4.2.1. Fine-tuning Performance

To evaluate the performance of EventDAM, we conduct experiments on the EventScape and MVSEC datasets. *Importantly, EventDAM’s training process does not require depth annotations, whereas the compared methods necessitate ground-truth depth.* On the EventScape dataset, we train and test EventDAM using the training and test sets. For the MVSEC dataset, we train EventDAM on the “Day2” subset and evaluate it on the “Day1”, “Night1”, “Night2”, and “Night3” subsets.

Comparison on EventScape dataset. In Tab. 1, we compare with SOTA methods on the EventScape dataset. In general, our EventDAM-S outperforms existing methods across most metrics. In addition, our EventDAM-L outperforms previous methods across all metrics. For example, with only event data as input, E2Depth [16] obtains a depth error of 1.79 at 10m, while EventDAM-S obtains a depth error of 0.59. For methods utilizing both event data and RGB images, SRFNet [30] achieves a depth error of 2.76 meters at 30m. In contrast, our EventDAM-L with only event input achieves an error of 2.30 meters, improving about **16.7%**. These results affirm that EventDAM successfully distills knowledge from DAM v2 [41] and adapts well to sparse event data, thus predicting accurate depth predictions compared to previous data-specific methods. The qualitative results in Fig. 5 demonstrate that EventDAM predicts clear structural details.

Comparison on MVSEC dataset. In Tab. 2, we further evaluate the performance of EventDAM on the MVSEC dataset. The results reveal that the proposed EventDAM significantly outperforms methods that utilize only event data, such as E2Depth [16]. For instance, at the 10m threshold in the “Night1” scene, the depth error for E2Depth [16] is 3.38 meters, whereas our EventDAM-L achieves a markedly lower error of 1.39 meters. However, our model does not outperform methods that integrate both event and image data, though it does surpass the SOTA method SRFNet [30] on certain metrics. The observation may be attributed to that our training process relies on the predictions of DAM. As shown in Fig. 7, the performance of DAM, when using grayscale image data on specific scenes, is suboptimal, which consequently leads to a decline in the performance of EventDAM.

4.2.2. Zero-shot Capability

To further assess the zero-shot capability of our EventDAM, we first train it on the EventScape dataset and subsequently assess its performance on the unseen DENSE and MVSEC datasets. *Note that the compared methods, such as RAMNet [11] and SRFNet [30], require ground-truth depth for model training on the DENSE dataset.*

DENSE dataset. We investigate the zero-shot capability of our method using the DENSE dataset. The results, presented in Tab. 3, demonstrate that our methods consistently outperforms existing SOTA approaches. Notably, our proposed EventDAM demonstrates superior performance compared to all other methods that utilize both event and image data. In particular, our EventDAM-L achieves 3.451 depth error at the 30-meter threshold, compared to 19.113 meters for RAMNet [11] and 6.116 meters for SRFNet [30]. These findings highlight the excellent zero-shot performance of our EventDAM for event-based depth estimation.

MVSEC dataset. To further evaluate the zero-shot capability of our proposed EventDAM, we conduct a fair comparison with DAM [41] on the unseen MVSEC “Night1” dataset, as presented in Tab. 4 and Fig. 6. The results demonstrate

Methods	Modality	Night1			Night2			Night3			Nay1		
		10m	20m	30m	10m	20m	30m	10m	20m	30m	10m	20m	30m
E2Depth [16]	E	3.38	3.82	4.46	1.67	2.63	3.58	1.42	2.33	3.18	1.67	2.64	3.13
RAMNet [11]	E+I	2.50	3.19	3.82	1.21	2.31	3.28	1.01	2.34	3.43	1.39	2.17	2.76
EvT ⁺ [36]		1.45	2.10	2.88	1.48	2.13	2.90	1.38	2.03	2.77	1.24	1.91	2.36
HMNet [14]		1.50	2.48	3.19	1.36	2.25	2.96	1.27	2.17	2.86	1.22	2.21	2.68
ER-F2D [6]		1.58	2.24	2.78	1.54	2.23	2.95	1.24	1.96	2.81	1.34	2.25	2.62
SRFNet [30]		1.26	1.95	3.01	1.19	2.13	3.22	1.01	2.12	3.52	0.96	1.77	2.37
EventDAM-S	E	1.38	2.54	3.04	1.44	2.13	3.09	1.50	2.17	3.15	1.12	1.73	2.49
EventDAM-B		1.39	1.98	3.19	1.45	2.21	3.19	1.48	2.17	3.15	1.07	1.70	2.49
EventDAM-L		1.39	2.10	3.25	1.43	2.18	3.22	1.44	2.16	3.22	1.12	1.79	2.69

Table 2. Comparison with methods across varying distances and sub-datasets on MVSEC, evaluated by absolute mean depth error.

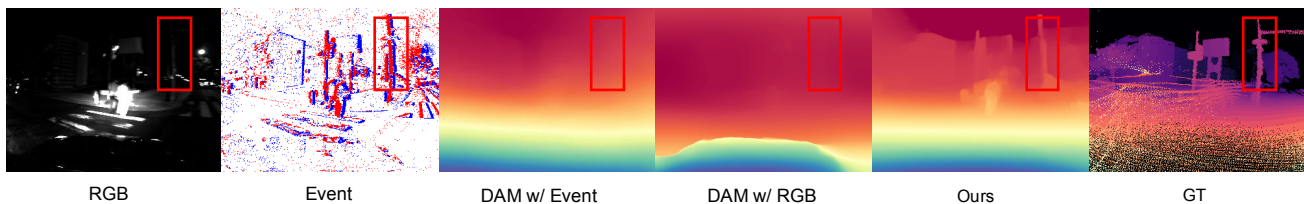


Figure 6. Qualitative results on MVSEC Night1 dataset.

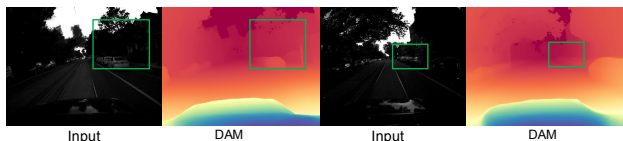


Figure 7. Qualitative results of DAM on MVSEC Day2.

that EventDAM consistently outperforms DAM [41] when relying exclusively on event data. For example, at a 30-meter distance, the depth error for DAM is 4.25, compared to 3.36 for EventDAM. Conversely, when image data serves as the input, DAM-S performs better than EventDAM-S at distances of 10m, 20m, and 30m. However, EventDAM-S achieves higher accuracy than DAM-S across the δ_1 , δ_2 , and δ_3 thresholds. The results show that our EventDAM with event data as input obtains depth estimation results that are slightly inferior to DAM with image input. It reveals that our proposed distillation strategies successively preserve the strong zero-shot capabilities of DAM, while improving its representation capabilities for the challenging event data that have distinct differences with conventional RGB image data.

4.3. Ablation Study

For all ablation studies, we utilize the EventScape dataset in conjunction with the DAM (ViT-S).

Effect of each component. To evaluate the contribution of each component within EventDAM, we conduct an ablation study on EventScape, as detailed in Tab. 5. The baseline model, with all components disabled and no knowledge distillation performed, exhibits relatively high errors of (1.35, 3.34, 4.92) at 10m, 20m, and 30m, respectively. Enabling \mathcal{L}_{SFM} reduces errors to (0.62, 1.75, 2.73), while using \mathcal{L}_{SFD} yields errors to (0.70, 1.79, 2.76). Additionally,

Method	Input	10m ↓	20m ↓	30m ↓
RAMNet [11]	E+I	2.619	11.264	19.113
SRFNet [30]		1.503	3.566	6.116
EventDAM-S	E	1.202	2.603	5.178
EventDAM-B		0.305	1.711	3.840
EventDAM-L		0.270	1.600	3.451

Table 3. Absolute mean depth error results on DENSE (in meters).

Method	Input	10m ↓	20m ↓	30m ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DAM-S [41]	E	2.15	3.22	4.25	0.21	0.36	0.45
DAM-S [41]	I	1.33	2.03	3.06	0.28	0.42	0.49
EventDAM-S	E	1.64	2.46	3.36	0.29	0.43	0.51

Table 4. Absolute mean depth error results on MVSEC Night1.

enabling \mathcal{L}_{SCM} also reduces errors to (0.67, 1.78, 2.72). Combining all three components achieves the best results, with errors of (0.59, 1.60, 2.47). These results demonstrate the effectiveness of the proposed components, which facilitate distillation from DAM to event-based DAM.

Effect of sparsity guidance for feature mixture. We investigate the effect of different feature mixing strategies in Tab. 6. Without feature mixing, the event-based DAM obtains depth errors of (0.65, 1.92, 2.99) at 10m, 20m, and 30m, respectively. Random feature mixing leads to slight performance degradation, with errors of (0.76, 2.15, 3.50), indicating that neglecting the sparsity might cause performance dropped. In contrast, sparsity-guided feature mixing significantly enhances model performance, reducing errors to (0.62, 1.75, 2.73), surpassing other mixing strategies. The results demonstrate that SFM enables dense-to-sparse distillation effectively.

Effect of sparsity guidance for feature distillation. We

\mathcal{L}_{SFM}	\mathcal{L}_{SFD}	\mathcal{L}_{SCM}	10m ↓	20m ↓	30m ↓
✗	✗	✗	1.35	3.34	4.92
✓	✗	✗	0.62	1.75	2.73
✗	✓	✗	0.70	1.79	2.76
✗	✗	✓	0.67	1.78	2.72
✓	✓	✗	0.60	1.64	2.57
✓	✓	✓	0.59	1.60	2.47

Table 5. Effect of each component of EventDAM on EventScape.

Feature Mixing Strategy	10m ↓	20m ↓	30m ↓
Without Mixing	0.65	1.92	2.99
Random Mixing	0.76	2.15	3.50
Mixing with Sparsity Guidance	0.62	1.75	2.73

Table 6. Ablation study about mixing features with or without sparsity guidance on EventScape.

Feature Distillation	10m ↓	20m ↓	30m ↓
Without Sparsity Guidance	0.79	2.01	3.00
With Sparsity Guidance	0.70	1.79	2.76

Table 7. Ablation study about feature distillation with or without sparsity guidance on EventScape.

conduct an ablation study to investigate the effect of sparsity guidance on feature distillation in EventScape. The results in Tab. 7 indicate that without sparsity guidance, EventDAM achieves errors of (0.79, 2.01, 3.00) at 10m, 20m, and 30m, respectively. When sparsity guidance is incorporated, EventDAM achieves errors of (0.70, 1.79, 2.76). These findings suggest that sparsity guidance plays a critical role in enhancing the dense-to-sparse distillation process, and that SFD facilitates DAM for event-based depth estimation.

4.4. Discussion

Fine-tuning layers and approaches. To comprehensively explore the potential of DAM in event-based depth estimation, we employ LoRA [18] to fine-tune the encoder while keeping the depth decoder fixed. To verify the effectiveness of this strategy, we conduct a series of ablation studies showing network performance with re-training different layers and different fine-tuning approaches. As shown in Tab. 8, we could observe an incremental performance improvement after fine-tuning DAM, which indicates that fine-tuning DAM could indeed improve the event-based depth estimation performance. Furthermore, by comparing the performance across different layers and fine-tuning approaches, we find that employing LoRA [18] on the encoder while keeping the decoder fixed yields substantial improvements in overall network performance.

Computational complexity. With only LoRA added, parameters of EventDAM are similar to DAM v2 [41]. As for the inference time, processing a 252×504 event data requires 12/22/58ms with EventDAM- $\{S,B,L\}$, respectively. The inference time is tested on an A100 GPU.

Fine-tuning Strategy	10m ↓	20m ↓	30m ↓
Fixed Model	1.35	3.34	4.92
Fine-tune All Layers	0.80	2.27	3.33
Fine-tune MLPs	0.81	2.20	3.11
LoRA Encoder + Fine-tune Decoder	0.69	1.87	2.88
LoRA Encoder + Fixed Decoder	0.60	1.64	2.57

Table 8. Ablation study about fine-tuning strategies with EventDAM on EventScape dataset.

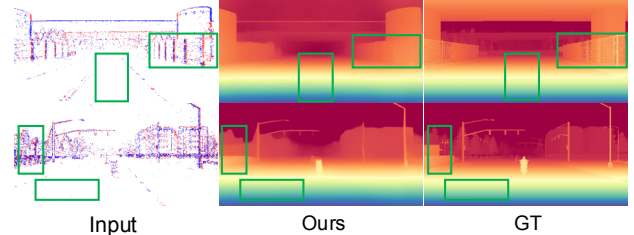


Figure 8. Qualitative results of EventDAM on EventScape.

Failure case. The visual results presented in Fig. 8 highlight that EventDAM struggles to effectively manage event-free regions and maintain the high-frequency details of small objects. These challenges are primarily due to the inherent sparsity of event data in these areas.

5. Conclusion and Limitation

In this study, we introduced EventDAM, a *pioneering approach* that leverages dense-to-sparse distillation to enhance event-based depth estimation without the need for annotated event data. By explicitly accounting for the inherent sparsity of event data, we propose three core components: Sparsity-aware Feature Mixture (SFM), Sparsity-aware Feature Distillation (SFD), and Sparsity-invariant Consistency Module (SCM). Through extensive experiments, we demonstrate the effectiveness of EventDAM, highlighting its superior performance and zero-shot capability across a variety of scenes. We hope this work will shed light on the future development of more scalable event-based depth estimation.

Limitation and future work: This study faces certain limitations due to the intrinsic characteristics of event cameras, particularly for depth predictions within event-free regions. Future research could focus on enhancing the restoration of these regions or integrating image-based depth estimation techniques for more effective solutions. Additionally, efforts will be directed toward acquiring more accurate depth datasets in complex environments and addressing the limitations of the DAM model under challenging conditions.

Acknowledgement: This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Education Bureau of Guangzhou.

References

- [1] Sk Miraj Ahmed, Suhas Lohit, Kuan-Chuan Peng, Michael J Jones, and Amit K Roy-Chowdhury. Cross-modal knowledge transfer without task-relevant source data. In *European Conference on Computer Vision*, pages 111–127. Springer, 2022. 3
- [2] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3
- [4] Yucheng Chen and Lin Wang. emoe-tracker: Environmental moe-based transformer for robust event-guided object tracking. *arXiv preprint arXiv:2406.20024*, 2024. 2
- [5] Zhiwen Chen, Zhiyu Zhu, Yifan Zhang, Junhui Hou, Guangming Shi, and Jinjian Wu. Segment any event streams via weighted adaptation of pivotal tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3890–3900, 2024. 2, 3
- [6] Anusha Devulapally, Md Fahim Faysal Khan, Siddharth Advani, and Vijaykrishnan Narayanan. Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2081–2089, 2024. 2, 5, 6, 7
- [7] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1
- [8] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018. 2
- [9] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5633–5643, 2019. 4
- [10] 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*, 6:2822–2829, 2021. 2, 4, 5
- [11] 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, 6, 7
- [12] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 2
- [13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 3
- [14] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22867–22876, 2023. 2, 5, 6, 7
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3
- [16] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. *2020 International Conference on 3D Vision (3DV)*, pages 534–542, 2020. 2, 4, 5, 6, 7
- [17] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4, 8
- [19] Sujin Jang, Dae Ung Jo, Sung Ju Hwang, Dongwook Lee, and Daehyun Ji. Stxd: structural and temporal cross-modal distillation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [20] Zexi Jia, Kaichao You, Weihua He, Yang Tian, Yongxiang Feng, Yaoyuan Wang, Xu Jia, Yihang Lou, Jingyi Zhang, Guoqi Li, et al. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32:1829–1842, 2023. 2
- [21] Linglin Jing, Yiming Ding, Yunpeng Gao, Zhigang Wang, Xu Yan, Dong Wang, Gerald Schaefer, Hui Fang, Bin Zhao, and Xuelong Li. Hpl-ess: Hybrid pseudo-labeling for unsupervised event-based semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23128–23137, 2024. 2
- [22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3
- [23] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15686–15698, 2024. 2, 3
- [24] Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2373–2383, 2023. 3
- [25] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. 2

- [26] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10016–10025, 2024. 4
- [27] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23–33, 2024. 1
- [28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 3
- [30] Tianbo Pan, Zidong Cao, and Lin Wang. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10695–10702. IEEE, 2024. 2, 4, 5, 6, 7
- [31] Yansong Peng, Yueyi Zhang, Peilin Xiao, Xiaoyan Sun, and Feng Wu. Better and faster: Adaptive event conversion for event-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2056–2064, 2023. 2
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3
- [33] Ulysse Rançon, Javier Cuadrado-Anibarro, Benoit R Cottereau, and Timothée Masquelier. Stereospike: Depth learning with a spiking neural network. *IEEE Access*, 10:127428–127439, 2022. 2
- [34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 4
- [35] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer. a sparse-aware solution for efficient event data processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2686, 2022. 4
- [36] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer+. a multi-purpose solution for efficient event data processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 6, 7
- [37] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1–7. IEEE, 2023. 2
- [38] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019. 2
- [39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 1
- [40] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021. 3
- [41] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 2, 3, 4, 5, 6, 7, 8
- [42] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13043–13052, 2021. 2
- [43] Xu Zheng and Lin Wang. Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17448–17458, 2024. 1
- [44] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023. 2
- [45] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay R. Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3: 2032–2039, 2018. 2, 5
- [46] Jinjing Zhu, Yunhao Luo, Xu Zheng, Hao Wang, and Lin Wang. A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11720–11730, 2023. 3
- [47] Jinjing Zhu, Yucheng Chen, and Lin Wang. Clip the divergence: language-guided unsupervised domain adaptation. *arXiv preprint arXiv:2407.01842*, 2024. 3
- [48] Jinjing Zhu, Yucheng Chen, and Lin Wang. Source-free cross-modal knowledge transfer by unleashing the potential of task-irrelevant data. *arXiv preprint arXiv:2401.05014*, 2024. 3