EventVGGT: Exploring Cross-Modal Distillation for Consistent Event-based Depth Estimation

Yinrui Ren[1,2*], Jinjing Zhu[1,3* †], Kanghao Chen[1*], Zhuoxiao Li[1*], Jing OU[1], Zidong Cao[1], Tongyan Hua[1], Peilun Shi[3], Yingchun Fu[2], Wufan Zhao[1†], and Hui Xiong[1†]

* Equal Contribution, † Project Lead, † Corresponding author

[1] Hong Kong University of Science and Technology (Guangzhou)
[2] South China Normal University
[3] Chinese University of Hong Kong

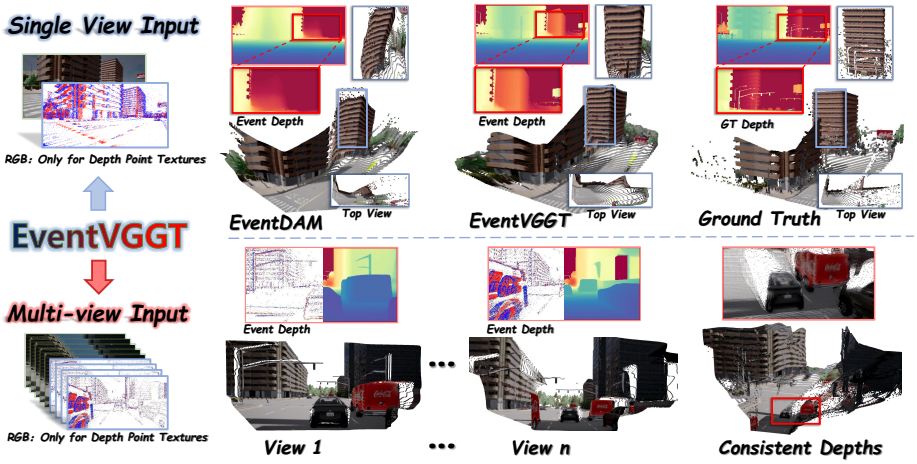

Fig. 1: We introduce **EventVGGT**, a novel framework that explicitly models asynchronous event streams as coherent video sequences. By distilling robust multi-view geometric priors from the Visual Geometry Grounded Transformer (VGGT), our approach achieves highly accurate and temporally consistent event-based depth estimation. *Note*: Depth inference relies exclusively on event data; RGB images are used solely for colorizing the reconstructed 3D point clouds.

Abstract. Event cameras offer superior sensitivity to high-speed motion and extreme lighting, making event-based monocular depth estimation a promising approach for robust 3D perception in challenging conditions. However, progress is severely hindered by the scarcity of dense depth annotations. While recent annotation-free approaches mitigate this by distilling knowledge from Vision Foundation Models (VFMs), a critical limitation persists: they process event streams as independent frames. By

neglecting the inherent temporal continuity of event data, these methods fail to leverage the rich temporal priors encoded in VFMs, ultimately yielding temporally inconsistent and less accurate depth predictions. To address this, we introduce **EventVGGT**, a novel framework that explicitly models the event stream as a coherent video sequence. To the best of our knowledge, we are the first to distill spatio-temporal and multi-view geometric priors from the Visual Geometry Grounded Transformer (VGGT) into the event domain. We achieve this via a comprehensive tri-level distillation strategy: (i) **Cross-Modal Feature Mixture (CMFM)** bridges the modality gap at the output level by fusing RGB and event features to generate auxiliary depth predictions; (ii) **Spatio-Temporal Feature Distillation (STFD)** distills VGGT's powerful spatio-temporal representations at the feature level; and (iii) **Temporal Consistency Distillation (TCD)** enforces cross-frame coherence at the temporal level by aligning inter-frame depth changes. Extensive experiments demonstrate that EventVGGT consistently outperforms existing methods—reducing the absolute mean depth error at 30m by over 53% on EventScape (from 2.30 to 1.06)—while exhibiting robust zero-shot generalization on the unseen DENSE and MVSEC datasets. The code is available at `https://github.com/yinruiRen/EventVGGT`.

Keywords: Depth Estimation · Event Camera · Vision Foundation Models · Cross-Modal Knowledge Distillation

1 Introduction

Event cameras [9, 45] are bio-inspired sensors that asynchronously encode logarithmic intensity changes into sparse event streams. Unlike conventional RGB cameras, they offer exceptional temporal resolution and dynamic range, enabling robust perception under high-speed motion and challenging illumination conditions [25, 39]. Consequently, they have driven significant advancements across various vision tasks, including object detection [24, 31], tracking [6, 10, 44], and segmentation [7, 19, 20, 22]. Among these, event-based monocular depth estimation [8, 16, 30] has emerged as a critical capability for autonomous driving and robotic navigation [33, 40]. However, despite recent efforts to fuse the complementary strengths of event and RGB modalities [14, 30], progress in event-based depth estimation remains bottlenecked by the scarcity of large-scale datasets with dense depth annotations.

To address this scarcity, recent methods [2, 50] leverage Vision Foundation Models (VFMs) to generate high-quality pseudo-depth labels from paired RGB-event data, bypassing the need for ground-truth annotations. For instance, EventDAM [50] employs dense-to-sparse distillation on isolated event frames, while DepthAnyEvent [2] integrates temporal modules before transferring knowledge from a single-image teacher (DAM V2 [43]). Although these VFM-based approaches alleviate the annotation bottleneck, a critical limitation persists: *by treating event streams as independent frames, they fail to leverage the inherent*

temporal continuity of the data and the temporal priors of VFMs, ultimately yielding temporally inconsistent and suboptimal depth predictions.

Among recent VFMs, models equipped with multi-view geometric reasoning are ideal teachers, as their spatio-temporal 3D priors naturally align with the continuous dynamics of event streams. A prime example is the Visual Geometry Grounded Transformer (VGGT) [41], which jointly infers depth, camera poses, and point clouds from multiple views. Unlike single-image baselines (*e.g.*, DAM V2 [43]) that lack cross-frame awareness, VGGT implicitly enforces geometric consistency across views via alternating frame-wise and global attention. Inspired by this, we propose to adapt VGGT's powerful multi-view reasoning for consistent event-based depth estimation. By treating the asynchronous event stream as a dense sequence of views, we explicitly exploit VGGT to model the temporal constraints and geometric coherence inherent in event data.

To this end, we introduce **EventVGGT**, an annotation-free framework that treats the event stream as a continuous video sequence and distills rich spatio-temporal priors from a VGGT teacher to achieve consistent event-based depth estimation (see Fig. 1). To effectively transfer this knowledge, we propose a tri-level distillation strategy: (i) At the *output level*, we introduce the **Cross-Modal Feature Mixture (CMFM)** module, which bridges the substantial modality gap by fusing RGB and event features to generate an auxiliary depth prediction. By explicitly supervising this auxiliary output with the teacher's high-fidelity RGB depth maps, we effectively distill the teacher's geometric priors into the event-based student. (ii) At the *feature level*, to fully exploit the teacher's internal representations, we propose **Spatio-Temporal Feature Distillation (STFD)**, aligning the student's intermediate features with those of VGGT to enforce the learning of spatial structure and temporal correspondence. (iii) At the *temporal level*, we introduce a **Temporal Consistency Distillation (TCD)** loss to supervise the inter-frame changes of consecutive depth maps rather than their absolute values. By penalizing discrepancies in inter-frame depth changes, TCD compels the student to inherit the teacher's geometrically coherent temporal flow, yielding accurate and stable depth sequences.

We validate EventVGGT on EventScape [11] and MVSEC [46] datasets, achieving new state-of-the-art performance against both event- and image-based methods. Furthermore, we demonstrate its robust zero-shot generalization: when trained solely on EventScape and evaluated on the unseen DENSE [16] and MVSEC datasets, EventVGGT maintains a substantial performance margin. The versatility of our approach is further highlighted by its effective extension to distilling VGGT for other geometric tasks, including camera pose and point map estimation from events.

In summary, our main contributions are three-fold:

- We introduce EventVGGT, the first framework to distill spatio-temporal priors from a multi-view foundation model (VGGT) into an event-based student, enabling temporally consistent, annotation-free depth estimation.

- We propose a comprehensive tri-level distillation strategy: Cross-Modal Feature Mixture (CMFM), Spatio-Temporal Feature Distillation (STFD), and Temporal Consistency Distillation (TCD).
- EventVGGT achieves state-of-the-art results on EventScape and MVSEC, demonstrates strong zero-shot generalization on DENSE, and seamlessly extends to camera pose and point cloud estimation.

2 Related Work

Event-based monocular depth estimation. Traditional RGB-based monocular depth estimation [3, 4, 43] is constrained by fixed frame rates and degrades severely under challenging illumination. To overcome these limitations, event-based depth estimation has emerged as a robust alternative, enabling high-frequency predictions in dynamic lighting and rapid motion scenarios [37, 45]. Early works, such as E2Depth [16], leverage recurrent convolutional networks to generate dense depth maps from event streams. Subsequent methods [8, 12, 26, 30,36] integrate paired RGB and event data to further enhance accuracy. Despite these advancements, most supervised approaches remain bottlenecked by their reliance on expensive, extensively annotated datasets. To bypass this, recent annotation-free works like DepthAnyEvent [2] and EventDAM [50] utilize cross-modal distillation to transfer knowledge from image-based Vision Foundation Models (VFMs). However, by treating event streams as isolated frames, these methods fail to adequately exploit the inherent temporal dynamics of events and the temporal priors of VFMs. *To address this, our work explicitly models the event stream as a continuous video sequence, ensuring that the estimated depth maps maintain strict geometric consistency throughout the sequence.*

Cross-modal knowledge distillation. Knowledge distillation (KD) classically transfers learned representations from a high-capacity teacher to a more compact student network [17, 42, 47, 49]. Building upon this, cross-modal KD [18, 23] extends the paradigm to bridge distinct sensory domains. For instance, early approaches [13] facilitate representation learning in an unlabeled target modality by exploiting mid-level features from a paired, labeled source. More recently, methods [1, 48] have tackled the challenging setting of transferring knowledge without direct access to task-relevant source data. *In this study, rather than relying on standard feature matching, we propose a comprehensive tri-level distillation strategy (CMFM, STFD, and TCD) to effectively bridge the substantial modality gap between paired RGB sequences and asynchronous event streams.*

Vision foundation models (VFMs). VFMs have demonstrated remarkable generalization capabilities, driven by large-scale datasets [21,32,43] and advances in self-supervised learning [5, 15, 29]. For example, models like CLIP [32] and SAM [21, 34] exhibit robust zero-shot transfer across diverse downstream tasks, while the DINO family [29, 38] extracts highly versatile visual representations from massive curated corpora. Building on these breakthroughs, recent efforts have adapted VFMs to alternative modalities, harnessing their representational power for tasks like semantic segmentation [7, 22] and depth estimation [2, 50].

Specifically, EventDAM [50] employs a dense-to-sparse distillation strategy to bypass the need for dense depth annotations, while DepthAnyEvent [2] utilizes cross-modal distillation to transfer the rich visual priors of image-based VFMs into the event domain. *Advancing this paradigm, we introduce a framework that explicitly distills the rich multi-view 3D priors from a sequence-aware foundation model (VGGT), effectively elevating event-based distillation from the spatial to the spatio-temporal domain.*

3 Method

3.1 Overview

Our study represents an initial effort to leverage VGGT [41] to achieve consistent event-based depth estimation, operating entirely without requiring access to ground-truth depth data. Fig. 2 illustrates the overall architecture of EventVGGT. We first detail the foundational components—including the inputs, event frame-like representation, and feature encoding—in Sec. 3.2. To facilitate robust knowledge transfer from the RGB-based VGGT teacher to the event-based EventVGGT student, we propose a tri-level distillation strategy comprising three core modules: Cross-Modal Feature Mixture (CMFM, Sec. 3.3), Spatio-Temporal Feature Distillation (STFD, Sec. 3.4), and Temporal Consistency Distillation (TCD, Sec. 3.5).

3.2 Event-based Monocular Depth Estimation

Inputs. Unlike conventional frame-based sensors that record absolute intensity values at fixed intervals, event cameras capture visual information as a continuous stream of asynchronous intensity changes. Formally, each event is defined as a tuple $e = (x, y, t, p)$, where (x, y) denotes the pixel coordinates, t is the precise timestamp of the change, and $p \in \{+1, -1\}$ indicates the polarity (representing an increase or decrease in logarithmic intensity). Concurrently, an RGB camera captures standard intensity images, denoted as $I^{img} \in \mathbb{R}^{3\times H\times W}$, where H and W are the spatial dimensions. Ultimately, the input to the EventVGGT framework comprises a synchronized sequence of N image frames, $I^{img} = \{I_i^{img}\}_{i=1}^{N}$, alongside the corresponding continuous event stream.

Event frame-like representation. To facilitate efficient computation and ensure compatibility with standard vision architectures, the continuous event stream is partitioned into fixed temporal windows and accumulated into dense frame-like representations. Following prior studies [11, 16, 30, 35], we divide the event stream into discrete temporal windows of $\Delta T = 50$ ms, discretizing each window into $B = 5$ temporal bins. Positive and negative polarity events are subsequently aggregated within each bin to construct histogram-based channels, yielding a synchronized event sequence $I^{\text{evt}} = \{I_i^{\text{evt}}\}_{i=1}^{N}$, where each frame $I_i^{\text{evt}} \in \mathbb{R}^{3\times H\times W}$. Finally, these event representations are temporally synchronized with the RGB images, producing perfectly aligned input sequences for the EventVGGT network.

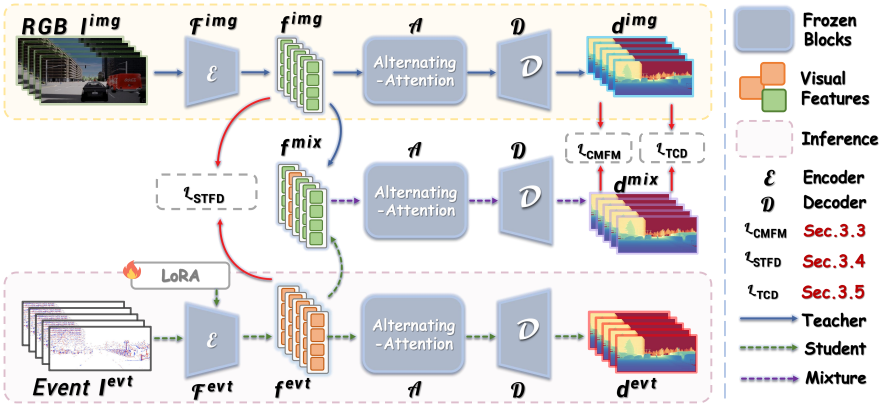

Fig. 2: Overview of the EventVGGT framework. Our approach distills robust multi-view geometric priors from VGGT to achieve consistent event-based depth estimation. Given synchronized event sequences I^{evt} and RGB images I^{img}, we extract modality-specific features using an event encoder $\mathcal{F}^{evt}$ and an image encoder $\mathcal{F}^{img}$, respectively. We introduce three core modules: (i) **Cross-Modal Feature Mixture (CMFM)**, which constructs mixed feature representations to predict auxiliary depths d^{mix} (*cf.* Sec. 3.3); (ii) **Spatio-Temporal Feature Distillation (STFD)**, which aligns intra-frame spatial and inter-frame temporal representations. (*cf.* Sec. 3.4); and (iii) **Temporal Consistency Distillation (TCD)**, which explicitly enforces geometric stability across consecutive depth predictions (*cf.* Sec. 3.5).

Feature encoding. Each frame in the RGB sequence I^{img} and the event sequence I^{evt} is divided into L non-overlapping square patches of size $P \times P$, where $L = \frac{H \times W}{P^2}$. The event-based student encoder $\mathcal{F}^{evt} : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^{D\times \hat{H}\times \hat{W}}$ processes the event frame sequence to output D-dimensional feature maps $f^{evt} = \{f_i^{evt}\}_{i=1}^{N}$, with reduced spatial dimensions $\hat{H} = \frac{H}{P}$ and $\hat{W} = \frac{W}{P}$. Similarly, the RGB-based teacher encoder $\mathcal{F}^{img}$ produces image feature maps $f^{img} = \{f_i^{img}\}_{i=1}^{N}$ of identical dimensions. Both encoders share a core network architecture comprising an Alternating-Attention Transformer $\mathcal{A}$ from VGGT [41], followed by the depth prediction head. Finally, a depth decoder $\mathcal{D}$ predicts the depth maps $d^{evt} = \{d_i^{evt}\}_{i=1}^{N}$ and $d^{img} = \{d_i^{img}\}_{i=1}^{N}$ from the corresponding event and image features, respectively.

3.3 Cross-Modal Feature Mixture

Due to the substantial modality gap between dense, absolute-intensity RGB images and sparse, asynchronous event streams, directly forcing an event-based student to mimic an RGB teacher's output space often leads to severe gradient conflicts, unstable convergence, and suboptimal adaptation. To overcome this bottleneck, we introduce the Cross-Modal Feature Mixture (**CMFM**) module at *output level*. Rather than enforcing an overly rigid, direct supervision that hinders the training process, CMFM constructs a "stepping stone" to gently bridge

the modalities at the output level. By stochastically mixing RGB and event features into a unified sequence and decoding it into an auxiliary prediction, we compel the shared decoder to treat event representations as functionally equivalent to RGB features. This strategy naturally pulls the event stream into the teacher's highly structured feature space, facilitating a much smoother knowledge transfer.

Formally, given the extracted frame-level feature representations (f^{img} and f^{evt}) from the synchronized RGB sequence I^{img} and event sequence I^{evt}, we synthesize a mixed feature sequence f^{mix}. This is achieved by stochastically substituting a subset of the RGB features with their temporally aligned event counterparts. Empirically, we uniformly randomly replace 25% of the RGB features with event features (e.g., resulting in a sequence like $f^{mix} = \{f_1^{img}, f_2^{evt}, f_3^{img}, \ldots, f_N^{img}\}$). This specific substitution ratio strikes an optimal balance, providing enough RGB context to maintain geometric stability while introducing sufficient event features to enforce cross-modal alignment (comprehensive justification is provided in Tab. 7).

Once constructed, the mixed feature sequence f^{mix} is fed through the shared Alternating-Attention Transformer $\mathcal{A}$ and depth decoder $\mathcal{D}$ to generate an auxiliary sequence of depth maps, $d^{mix} = \{d_i^{mix}\}_{i=1}^N$. By explicitly supervising this auxiliary output with the high-fidelity depth d^{img} generated by the VGGT teacher, the distillation process is substantially stabilized.

Following the optimization strategy of VGGT [41], our distillation loss incorporates an aleatoric-uncertainty formulation [28] to intelligently weight the discrepancy between d^{mix} and d^{img} based on the predicted confidence map $\Sigma^{d^{mix}}$. Furthermore, to explicitly preserve local structural consistency—a critical factor for dense monocular depth estimation—we augment this objective with a gradient-based penalty. The objective of CMFM module is formulated as:

$$\mathcal{L}_{\text{CMFM}} = \sum_{i=1}^{N} \Big(\|\Sigma_i^{d^{img}} \odot (d_i^{mix} - d_i^{img})\| + \|\Sigma_i^{d^{img}} \odot (\nabla d_i^{mix} - \nabla d_i^{img})\| - \alpha \log \Sigma_i^{d^{img}} \Big), \tag{1}$$

where $\odot$ denotes the channel-broadcast element-wise product and α is a balancing hyperparameter. Conceptually, $\mathcal{L}_{\text{CMFM}}$ serves as a vital cross-modal bridge. By aligning the mixed-feature auxiliary predictions with the teacher's outputs, it not only stabilizes training but ensures the student model successfully inherits the teacher's robust geometric priors for event-based depth estimation.

3.4 Spatio-Temporal Feature Distillation

While CMFM successfully bridges the modality gap at the output level, the student network must also inherit the rich geometric priors embedded deep within the VGGT teacher's intermediate representations. However, this presents a fundamental challenge: *event streams intrinsically encode ultra-high-frequency temporal dynamics (motion), whereas the VGGT teacher is optimized to capture spatial scene structures across discrete, static frames.* Prior works, such as

EventDAM [50] and DepthAnyEvent [2], ignore this discrepancy by processing event streams as independent static frames. This standard frame-by-frame alignment forces dynamic event features into a rigid static representation, destroying their inherent temporal structure and leading to temporally inconsistent and less accurate depth predictions.

To address this discrepancy, we introduce Spatio-Temporal Feature Distillation (**STFD**) at *feature level*. Rather than merely enforcing a rigid spatial mimicry that suppresses event dynamics, STFD explicitly models and transfers the temporal changes of the features. Formally, given the RGB features f^{img} and corresponding event features f^{evt}, our objective is composed of two complementary terms:

$$\mathcal{L}_{\text{STFD}} = \sum_{i=1}^{N} \big(1 - \cos(f_i^{evt}, f_i^{img})\big) + \sum_{i=1}^{N-1} \big(1 - \cos(f_{i+1}^{evt} - f_i^{evt}, f_{i+1}^{img} - f_i^{img})\big), \quad (2)$$

where $\cos(\cdot)$ denotes the channel-wise cosine similarity. The first term performs intra-frame spatial distillation, anchoring the event features to the robust structural geometry embedded within the VGGT backbone. Crucially, the second term executes inter-frame temporal distillation by comparing frame-to-frame feature variations. By explicitly matching the magnitudes and directions of these temporal transitions in the latent space, we ensure that the student network learns motion-sensitive dynamics that are consistent with the RGB teacher's temporal reasoning, fully exploiting the continuous nature of the event stream.

3.5 Temporal Consistency Distillation

While CMFM and STFD successfully align the modalities at the spatial output and intermediate feature levels, a significant challenge remains at the final prediction stage. Because event streams capture sparse, asynchronous brightness changes rather than dense photometric data, event-based dense predictions are inherently susceptible to high-frequency temporal instability, often manifesting as severe depth flickering. Standard distillation objectives fail to address this, as they naively penalize per-frame absolute depth errors while ignoring inherent temporal continuity and geometric coherence across frames.

To close this gap, we introduce Temporal Consistency Distillation (**TCD**) at *temporal level*. The teacher model, VGGT, inherently enforces multi-view geometric consistency across all frames simultaneously via its global feed-forward architecture, producing depth sequences with dynamically smooth, physically realistic transitions. *Our key innovation is to explicitly align the student's inter-frame depth variations with the temporally coherent transitions of the teacher model*. We formulate the TCD objective to penalize discrepancies in the inter-frame rate of change between the student and teacher outputs:

$$\mathcal{L}_{\text{TCD}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \Big\| \, |d_{i+1}^{mix} - d_i^{mix}| - |d_{i+1}^{img} - d_i^{img}| \, \Big\|_1. \quad (3)$$

The term $|d_{i+1}^{img} - d_i^{img}|$ calculates the per-pixel magnitude of depth variation across consecutive frames, functioning as a continuous temporal gradient map. This gradient captures the scene's underlying geometric dynamics, including camera ego-motion and dynamic objects. By minimizing the L_1 distance between these temporal difference fields, $\mathcal{L}_{\text{TCD}}$ acts as a strict temporal-geometric constraint. It effectively suppresses physically implausible inter-frame discontinuities, anchoring the student's output dynamics to the multi-view-optimized evolution of the VGGT teacher. Ultimately, this guarantees that the student produces spatially precise and dynamically coherent depth sequences.

Total objective. Combining our proposed distillation components, the final training objective of the EventVGGT framework is formulated as a weighted combination of cross-modal, spatio-temporal, and temporal consistency losses:

$$\mathcal{L} = \mathcal{L}_{\text{CMFM}} + \lambda_{\text{STFD}}\mathcal{L}_{\text{STFD}} + \lambda_{\text{TCD}}\mathcal{L}_{\text{TCD}}, \tag{4}$$

where λ_{STFD} and λ_{TCD} serve as balancing hyperparameters for the feature-level alignment and temporal consistency distillation, respectively. Empirically, we set $\lambda_{\text{STFD}} = 0.1$ and $\lambda_{\text{TCD}} = 0.2$ to effectively regularize the training process.

4 Experiment

4.1 Settings

Datasets. We adopt EventScape [12] as our primary synthetic training source due to its large-scale scenes and dense depth annotations. For standard in-domain evaluation, EventVGGT is trained and tested on the respective splits of EventScape and MVSEC [46]. To assess zero-shot generalization, we train exclusively on EventScape and evaluate directly on the real-world MVSEC and the unseen synthetic DENSE [16] datasets.

Evaluation metrics. Following prior works [8,14,30], we measure average absolute depth error at cut-off distances of 10m, 20m, and 30m. Furthermore, we evaluate our approach using three percentage metrics δ_i, where $i \in \{1.25, 1.25^2, 1.25^3\}$.

Implementation details. Consistent with prior works [30, 50], we center-crop both the RGB images and their corresponding event representations from the EventScape dataset to a spatial resolution of 252×504 pixels. Aligning with the VGGT training protocol [41], sky regions are masked out to exclude invalid depth values. We optimize the network using AdamW with a learning rate of 1×10^{-4}. For parameter-efficient fine-tuning [50], we apply LoRA ($r = 16$ and $\alpha = 32$), adding approximately 1.6 million trainable parameters. The model is trained on four NVIDIA A800 GPUs with dynamic batch sizes, converging in roughly 11 hours. To effectively distill the rich temporal dynamics from VGGT, we input sequences of 24 synchronized RGB and event frames per batch.

4.2 Experimental Results

Following EventDAM [50], we evaluate the supervised fine-tuning performance of EventVGGT on EventScape and MVSEC datasets. For the EventScape dataset,

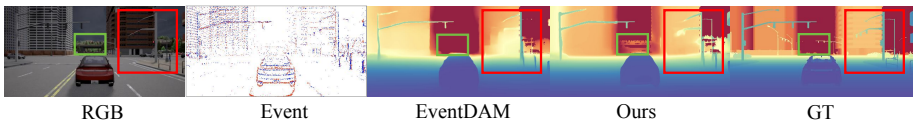

Fig. 3: Qualitative results on EventScape dataset.

the model is trained on training Town 1 subset and evaluated on test set. For MVSEC dataset, EventVGGT is trained on the "Day2" subset and subsequently evaluated across the "Day1", "Night1", "Night2", and "Night3" subsets.

Comparison on EventScape dataset. In Tab. 1, we compare the proposed EventVGGT against existing state-of-the-art methods on the EventScape dataset. Overall, EventVGGT consistently outperforms all baselines across evaluated metrics, with the most substantial performance gains observed at greater distances. For instance, while EventDAM yields a depth error of 2.30 at 30m, our method reduces this error to 1.06, representing a 53.9% relative improvement.

Furthermore, EventVGGT yields lower error rates compared to existing methods that require both event and RGB data (Input: E+I) during inference. For instance, while SRFNet and ER-F2D report depth errors of 2.76 and 2.81 at 30m, respectively, EventVGGT achieves a 1.06 error at the same distance using event data alone. These results demonstrate that EventVGGT learns the robust spatio-temporal representations from the VGGT teacher, effectively bridging the modality gap without requiring auxiliary RGB input at test time. As shown in Fig. 3, visual comparisons on EventScape further validate our quantitative findings. As highlighted by the red bounding boxes, EventVGGT outperforms the EventDAM baseline in recovering fine-grained spatial structures. Our method preserves sharp object boundaries and accurately captures the depth of thin, challenging structures (e.g., streetlights), whereas EventDAM suffers from geometric blurring. Conversely, the green boxes illustrate a specific characteristic of our model: it tends to slightly underestimate the depth of extremely distant background elements, such as far-off buildings. This phenomenon is a direct artifact of the cross-modal distillation process, as VGGT inherently exhibits a similar far-field depth compression bias on this dataset (See Fig. 4).

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**.

Method	Input	10m ↓	20m ↓	30m ↓
E2Depth [16]	E	1.79	5.35	8.31
RAMNet [12]	E+I	0.81	2.26	3.58
HMNet [14]		**0.55**	1.80	3.27
ER-F2D [8]		0.67	1.69	2.81
SRFNet [30]		1.27	1.68	2.76
EventDAM [50]	E	0.56	**1.52**	**2.30**
EventVGGT		**0.54**	**0.79**	**1.06**

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

Method	Input	Night1			Night2			Night3			Day1		
		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 [12]	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 [8]		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
EReFormer [27]	E	1.52	2.28	2.98	1.40	2.12	2.66	1.32	2.04	**2.68**	1.29	2.14	2.59
EventDAM [50]		**1.39**	2.10	3.25	1.43	2.18	3.22	1.44	2.16	3.22	**1.12**	1.79	2.69
Depth AnyEvent [2]		1.87	2.27	2.81	1.99	2.40	**2.86**	2.05	2.49	2.97	1.50	1.97	2.40
EventVGGT		1.67	**2.02**	**2.61**	1.64	**2.03**	**2.48**	1.74	2.15	**2.64**	**0.96**	**1.33**	**1.63**

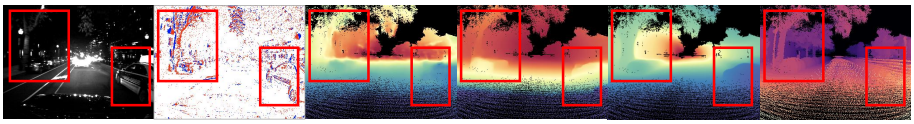

Fig. 5: Qualitative results on MVSEC Night1 dataset.

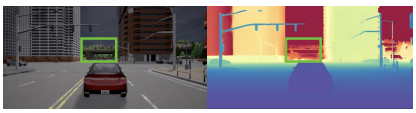

Fig. 4: Qualitative result of VGGT on EventScape dataset.

Comparison on MVSEC dataset. In Tab. 2, we evaluate EventVGGT on the real-world MVSEC dataset to assess its robustness against extreme illumination changes. Under these challenging conditions, EventVGGT consistently outperforms the event-only EventDAM, reducing the 30m absolute mean depth error from 3.22 to 2.48 on the Night 2 sequence, and from 3.22 to 2.64 on Night 3. While EventDAM's frame-wise distillation suffers from spatial ambiguities when processing erratic nighttime event distributions, EventVGGT overcomes this domain shift by utilizing STFD and TCD to model the continuous video sequence and stabilize predictions using cross-frame dynamics. Furthermore, EventVGGT achieves performance comparable to—and often exceeding—multi-modal (E+I) fusion approaches like SRFNet, which yields a higher 3.52 error at 30m on Night 3 despite relying on paired images. By leveraging the CMFM module during daytime training to deeply internalize multi-view geometric priors from the RGB-based VGGT teacher, our event-based student model effectively reconstructs missing spatial structures from pure event dynamics, bypassing the need for degraded RGB inputs during inference and mitigating the low-light failure modes of standard cameras. The visual results presented in Fig. 5 highlight the effectiveness of EventVGGT.

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

Method	Input	10m ↓	20m ↓	30m ↓
RAMNet [12]	E+I	2.62	11.26	19.11
SRFNet [30]		1.50	3.57	6.12
EventDAM [50]	E	**1.20**	**2.60**	**5.18**
EventVGGT		**0.54**	**0.89**	**1.33**

Table 4: Quantitative results on MVSEC Night1 dataset.

Method	Input	10m ↓	20m ↓	30m ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
VGGT [41]	E	2.42	2.71	3.44	**0.42**	0.69	0.84
VGGT [41]	I	**2.31**	**2.68**	**3.33**	0.41	**0.71**	**0.88**
EventVGGT	E	**1.67**	**2.02**	**2.61**	**0.57**	**0.82**	**0.93**

To further assess the zero-shot capability of our EventVGGT, we first train it on the EventScape dataset and subsequently assess its performance on the unseen DENSE and MVSEC datasets.

Comparison on DENSE dataset. In Tab. 3, we further investigate the zero-shot generalization capabilities of EventVGGT by training exclusively on the EventScape dataset and directly evaluating on the unseen DENSE dataset. The results highlight EventVGGT's exceptional capacity to transfer learned representations across distinct domains, vastly outperforming multi-modal baselines (Input: E+I) such as RAMNet and SRFNet. And EventVGGT (Input: E) restricts its error to a mere 1.33. Furthermore, EventVGGT substantially outperforms the event-only state-of-the-art EventDAM, which records a higher 5.18 error at 30m. Visual results on DENSE dataset are presented in Fig. 6.

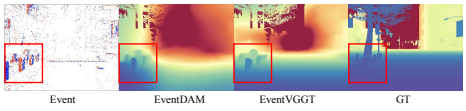

Fig. 6: Qualitative results on DENSE.

Comparison on MVSEC dataset. We perform a direct zero-shot comparison on the challenging MVSEC Night1 dataset to evaluate the effectiveness of our cross-modal distillation strategy against the original VGGT foundation model in Tab. 4. When the pre-trained VGGT is directly applied to event data (Input: E), it suffers from severe performance degradation due to the massive modality gap and density discrepancies between its native RGB training domain and sparse, asynchronous event streams. By contrast, EventVGGT (Input: E) significantly outperforms the raw VGGT on event data across all distance and δ accuracy metrics. More impressively, when comparing EventVGGT (Input: E) to the original VGGT operating on its native RGB images (Input: I), our method achieves highly competitive absolute depth errors and demonstrates superior robustness in the δ metrics. This demonstrates that our distillation pipeline does not merely mimic the teacher network, but effectively empowers the event-based student to leverage the high dynamic range of events in low-light night scenes where standard RGB images severely degrade.

4.3 Ablation Study

For all ablation studies, we utilize the training Town 1 dataset of EventScape to train EventVGGT and evaluate the performance with test dataset of EventScape.

Table 5: Effect of each component of EventVGGT on EventScape.

$\mathcal{L}_{\text{CMFM}}$	$\mathcal{L}_{\text{STFD}}$	$\mathcal{L}_{\text{TCD}}$	10m ↓	20m ↓	30m ↓
✗	✗	✗	0.69	0.94	1.26
✓	✗	✗	0.57	0.86	1.13
✓	✓	✗	0.53	0.81	1.10
✓	✗	✓	0.56	0.84	1.12
✓	✓	✓	0.51	0.79	1.06

Table 6: Ablation study on the input sequence length.

N	10m ↓	20m ↓	30m ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
1	0.79	1.10	1.50	0.82	0.96	0.99
4	0.75	1.05	1.41	0.83	0.97	0.99
8	0.70	0.98	1.31	0.85	0.97	0.99
24	0.71	0.95	1.26	0.85	0.97	0.99

Effect of each component. As detailed in Tab. 5, we validate the individual contributions of our core distillation modules. The naive baseline yields suboptimal depth errors of 0.69, 0.94, and 1.26 at the 10m, 20m, and 30m cut-offs, respectively. Integrating $\mathcal{L}_{\text{CMFM}}$ bridges the cross-modal gap, reducing these errors to 0.57, 0.86, and 1.13. Building on this, $\mathcal{L}_{\text{STFD}}$ explicitly aligns intra-frame geometry and inter-frame dynamics with the teacher model, further decreasing errors to 0.53, 0.81, and 1.10. Finally, the cumulative integration of $\mathcal{L}_{\text{TCD}}$ yields our full EventVGGT framework, achieving minimal errors of 0.51, 0.79, and 1.06. These results validate the necessity of integrating both feature-level alignment and prediction-level temporal consistency for robust cross-modal distillation.

Impact of input sequence length. In Tab. 6, we ablate the input sequence length to evaluate its impact on extracting spatio-temporal 3D priors from the VGGT teacher. Processing a single frame (N=1) yields suboptimal performance (e.g., a 30m error of 1.50), as the model lacks the multi-view context required for geometric reasoning. Incrementally increasing the sequence length to 4 and 8 frames significantly reduces depth errors and improves δ_1 accuracy. Extending the sequence to N=24 frames achieves the best long-range depth estimation, reaching minimum errors of 0.95 and 1.26 at the 20m and 30m cut-offs. These results support our design choice: processing extended 24-frame sequences fully exploits the temporal multi-view priors of VGGT, providing optimal supervision signals for our cross-modal distillation pipeline.

Table 7: Ablation study on the CMFM feature mixture rate.

Rate (%)	10m ↓	20m ↓	30m ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
100	0.70	0.95	1.26	0.85	0.97	**0.99**
75	0.68	0.94	1.26	0.86	0.97	**0.99**
50	0.60	0.87	1.15	**0.90**	**0.98**	**0.99**
25	**0.57**	**0.86**	**1.13**	**0.90**	**0.98**	**0.99**
10	0.60	0.88	1.16	0.89	**0.98**	**0.99**

Influence of CMFM mixture ratio. Tab. 7 reports the ablation study on the feature replacement rate within the $\mathcal{L}_{CMFM}$ module. Completely replacing RGB features with event features (100% rate) yields suboptimal depth estimation, as it removes the dense spatial anchors necessary for stable cross-modal distillation. Lowering the rate to 25% provides the optimal regularization balance, achieving minimal absolute depth errors (0.57, 0.86, and 1.13 at 10m, 20m, and 30m) and the highest δ_1 accuracy (0.90). This specific ratio preserves sufficient RGB structural context

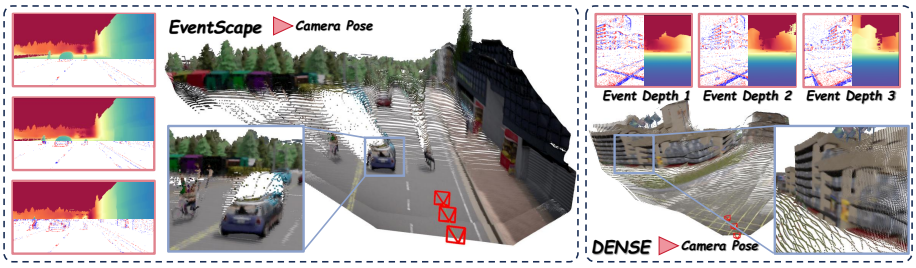

Fig. 7: Qualitative results on additional geometric estimation tasks.

Table 8: Ablation study on the impact of training data diversity.

Dataset	EventScape						DENSE					
	10m ↓	20m ↓	30m ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	10m ↓	20m ↓	30m ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Town 1	**0.51**	**0.79**	1.09	**0.92**	**0.98**	**0.99**	0.64	1.03	1.50	0.87	**0.98**	**0.99**
Town 2	0.61	0.86	1.16	0.89	**0.98**	**0.99**	0.79	1.17	1.72	0.79	0.96	**0.99**
Town 3	0.62	0.88	1.17	0.89	**0.98**	**0.99**	0.82	1.24	1.75	0.74	0.96	**0.99**
Towns 1-3	0.53	**0.79**	**1.06**	**0.92**	**0.98**	**0.99**	**0.54**	**0.89**	**1.33**	**0.87**	**0.98**	**0.99**

while effectively forcing the event encoder to integrate and compensate for the masked regions. Further reducing the rate to 10% degrades performance, as the sparse injection of event features weakens the distillation signal. Consequently, we adopt a 25% replacement rate empirically.

About training dataset. As reported in Tab. 8, training EventVGGT exclusively on the Town 1 subset yields in-domain performance on EventScape comparable to the model trained on the entire dataset (Towns 1-3). However, it suffers a significant performance drop during zero-shot evaluation on DENSE. This indicates that while expanding synthetic training diversity provides marginal in-domain gains, it is essential for robust cross-domain generalization.

4.4 Discussion

Extension to auxiliary 3D tasks. To demonstrate the versatility of our framework, we extend the distillation pipeline to facilitate camera pose and dense point cloud estimation. Qualitative results on EventScape and DENSE showcase the robust reconstruction of these broader geometric properties (See Fig. 7). This highlights the inherent flexibility of the EventVGGT, suggesting that our distilled spatio-temporal priors extend naturally to a broad spectrum of event-based 3D perception tasks beyond monocular depth estimation.

Qualitative analysis of depth consistency. To validate temporal stability, we back-project inferred multi-frame depths and poses into 3D point clouds. As shown in Fig. 7, EventVGGT seamlessly aggregates continuous event streams into unified reconstructions on EventScape and DENSE. By leveraging VGGT's multi-view priors, our framework eliminates the scale inconsistency of frame-by-frame methods, yielding accurate camera extrinsics (red frustums) and precise

geometric point cloud alignment. This high-fidelity reconstruction underscores our superiority in consistent event-based depth estimation.

Computational complexity. By employing LoRA for adaptation, the total parameters of EventVGGT remain highly comparable to that of VGGT [41]. Correspondingly, EventVGGT demonstrates high inference efficiency, processing a 252×504 event representation in 24 ms on a single NVIDIA A800 GPU.

5 Conclusion

We presented EventVGGT, a novel framework for temporally consistent, annotation-free event-based depth estimation. By explicitly modeling asynchronous event streams as continuous video sequences, our approach enables the distillation of robust spatio-temporal and multi-view geometric priors from the image-based VGGT to event domain. To effectively bridge the modality gap, we introduce a comprehensive cross-modal distillation strategy comprising Cross-Modal Feature Mixture (CMFM), Spatio-Temporal Feature Distillation (STFD), and Temporal Consistency Distillation (TCD). Extensive experiments demonstrate that EventVGGT establishes a new state-of-the-art on EventScape and MVSEC, exhibiting robust zero-shot generalization to unseen domains. Furthermore, our framework seamlessly extends to holistic 3D perception tasks, including camera pose and point cloud estimation.

Limitations and Future Work. EventVGGT inherits a minor far-field depth compression bias from the VGGT teacher model, occasionally leading to the underestimation of extremely distant backgrounds. Future work will integrate dense ground-truth depth into the distillation pipeline to explicitly calibrate this long-range artifact and enhance real-world reliability.

6 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 92370204), in part by the National Key R&D Program of China(Grant No. 2023YFF0725001), in part by Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence, in part by the guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities (Grant No. 2024ZDZX1007), in part by Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 42401567), in part by Guangdong Provincial Project (Grant No. 2024QN11G095), in part by The Open Fund of the State Key Laboratory of Spatial Datum (Grant No. SKLSD2026-KF-25), National Natural Science Foundation of China (Grant No. 42571390), in part by Science and Technology Projects of Xizang Autonomous Region, China(Grant No. XZ202501ZY0091), in part by Basic and Applied Basic Research Foundation of Guangdong Province (Grant No. 2025A1515011807), and in part by the Research Grants Council (RGC) of Hong Kong SAR (Grant Nos. GRF14213125, GRF14201824, and GRF14216222).

References

1. Ahmed, S.M., Lohit, S., Peng, K.C., Jones, M.J., Roy-Chowdhury, A.K.: Cross-modal knowledge transfer without task-relevant source data. In: European Conference on Computer Vision. pp. 111–127. Springer (2022)
2. Bartolomei, L., Mannocci, E., Tosi, F., Poggi, M., Mattoccia, S.: Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19669–19678 (2025)
3. Cao, Z., Zhu, J., Ai, H., Jiang, L., Lyu, Y., Xiong, H.: St ^{2}360 d: Spatial-to-temporal consistency for training-free 360 monocular depth estimation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
4. Cao, Z., Zhu, J., Zhang, W., Ai, H., Bai, H., Zhao, H., Wang, L.: Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 982–992 (2025)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Chen, Y., Wang, L.: emoe-tracker: Environmental moe-based transformer for robust event-guided object tracking. arXiv preprint arXiv:2406.20024 (2024)
7. Chen, Z., Zhu, Z., Zhang, Y., Hou, J., Shi, G., Wu, J.: Segment any event streams via weighted adaptation of pivotal tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3890–3900 (2024)
8. Devulapally, A., Khan, M.F.F., Advani, S., Narayanan, V.: 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. pp. 2081–2089 (2024)
9. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
10. Gehrig, D., Rebecq, H., Gallego, G., Scaramuzza, D.: Asynchronous, photometric feature tracking using events and frames. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 750–765 (2018)
11. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotics and Automation Letters **6**, 2822–2829 (2021), `https://api.semanticscholar.org/CorpusID:231951439`
12. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotics and Automation Letters **6**(2), 2822–2829 (2021)
13. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2827–2836 (2016)
14. Hamaguchi, R., Furukawa, Y., Onishi, M., Sakurada, K.: Hierarchical neural memory network for low latency event processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22867–22876 (2023)

15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
16. Hidalgo-Carri'o, J., Gehrig, D., Scaramuzza, D.: Learning monocular dense depth from events. 2020 International Conference on 3D Vision (3DV) pp. 534–542 (2020), https://api.semanticscholar.org/CorpusID:223957202
17. Hinton, G.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
18. Jang, S., Jo, D.U., Hwang, S.J., Lee, D., Ji, D.: Stxd: structural and temporal cross-modal distillation for multi-view 3d object detection. Advances in Neural Information Processing Systems **36** (2024)
19. Jia, Z., You, K., He, W., Tian, Y., Feng, Y., Wang, Y., Jia, X., Lou, Y., Zhang, J., Li, G., et al.: Event-based semantic segmentation with posterior attention. IEEE Transactions on Image Processing **32**, 1829–1842 (2023)
20. Jing, L., Ding, Y., Gao, Y., Wang, Z., Yan, X., Wang, D., Schaefer, G., Fang, H., Zhao, B., Li, X.: Hpl-ess: Hybrid pseudo-labeling for unsupervised event-based semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23128–23137 (2024)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
22. Kong, L., Liu, Y., Ng, L.X., Cottereau, B.R., Ooi, W.T.: Openess: Event-based semantic scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15686–15698 (2024)
23. Lee, P., Kim, T., Shim, M., Wee, D., Byun, H.: Decomposed cross-modal distillation for rgb-based temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2373–2383 (2023)
24. Li, J., Li, J., Zhu, L., Xiang, X., Huang, T., Tian, Y.: Asynchronous spatio-temporal memory network for continuous event-based object detection. IEEE Transactions on Image Processing **31**, 2975–2987 (2022)
25. Liang, G., Chen, K., Li, H., Lu, Y., Wang, L.: 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. pp. 23–33 (2024)
26. Liu, X., Fan, X., Li, J., Li, D., Zhang, W., Ma, Z., Tian, Y.: High-rate monocular depth estimation via cross frame-rate collaboration of frames and events. International Journal of Computer Vision **133**(10), 7332–7351 (2025)
27. Liu, X., Li, J., Shi, J., Fan, X., Tian, Y., Zhao, D.: Event-based monocular depth estimation with recurrent transformers. IEEE Trans. Cir. and Sys. for Video Technol. **34**(8), 7417–7429 (Aug 2024). https://doi.org/10.1109/TCSVT.2024.3378742, https://doi.org/10.1109/TCSVT.2024.3378742
28. Novotny, D., Larlus, D., Vedaldi, A.: Learning 3d object categories by looking around them. In: Proceedings of the IEEE international conference on computer vision. pp. 5218–5227 (2017)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Di-

nov2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023), https://api.semanticscholar.org/CorpusID:258170077
30. Pan, T., Cao, Z., Wang, L.: 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). pp. 10695–10702. IEEE (2024)
31. Peng, Y., Zhang, Y., Xiao, P., Sun, X., Wu, F.: Better and faster: Adaptive event conversion for event-based object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2056–2064 (2023)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
33. Rançon, U., Cuadrado-Anibarro, J., Cottereau, B.R., Masquelier, T.: Stereospike: Depth learning with a spiking neural network. IEEE Access **10**, 127428–127439 (2022)
34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
35. Sabater, A., Montesano, L., Murillo, A.C.: Event transformer. a sparse-aware solution for efficient event data processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2677–2686 (2022)
36. Sabater, A., Montesano, L., Murillo, A.C.: Event transformer+. a multi-purpose solution for efficient event data processing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
37. Shi, P., Peng, J., Qiu, J., Ju, X., Lo, F.P.W., Lo, B.: Even: An event-based framework for monocular depth estimation at adverse night conditions. In: 2023 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 1–7. IEEE (2023)
38. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
39. Tulyakov, S., Bochicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D.: 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. pp. 17755–17764 (2022)
40. Tulyakov, S., Fleuret, F., Kiefel, M., Gehler, P., Hirsch, M.: Learning an event sequence embedding for dense event-based deep stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1527–1537 (2019)
41. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
42. Wang, L., Yoon, K.J.: 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)
43. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
44. Zhang, J., Yang, X., Fu, Y., Wei, X., Yin, B., Dong, B.: Object tracking by jointly exploiting frame and event domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13043–13052 (2021)

45. Zheng, X., Liu, Y., Lu, Y., Hua, T., Pan, T., Zhang, W., Tao, D., Wang, L.: Deep learning for event-based vision: A comprehensive survey and benchmarks. arXiv preprint arXiv:2302.08890 (2023)
46. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V.R., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**, 2032–2039 (2018), `https://api.semanticscholar.org/CorpusID:3416874`
47. Zhu, J., Chen, Y., Wang, L.: Clip the divergence: language-guided unsupervised domain adaptation. arXiv preprint arXiv:2407.01842 (2024)
48. Zhu, J., Chen, Y., Wang, L.: Source-free cross-modal knowledge transfer by unleashing the potential of task-irrelevant data. arXiv preprint arXiv:2401.05014 (2024)
49. Zhu, J., Luo, Y., Zheng, X., Wang, H., Wang, L.: 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. pp. 11720–11730 (2023)
50. Zhu, J., Pan, T., Cao, Z., Liu, Y., Kwok, J.T., Xiong, H.: Depth any event stream: Enhancing event-based monocular depth estimation via dense-to-sparse distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5146–5155 (2025)