

Comparing Representations for Event Camera-based Visual Object Tracking

Oussama Abdul Hay, Sara Alansari, Mohamad Alansari, and Yahya Zweiri
Khalifa University of Science and Technology

Abstract

Visual object tracking (VOT) in dynamic environments is challenging due to Motion Blur (MB), Illumination Variations (IV), and Fast Motion (FM), conditions where traditional RGB-based trackers often fail. Event cameras offer high temporal resolution and low latency, making them well-suited for such scenarios. However, current event-based tracking methods rely on arbitrarily selected event representations, lacking systematic evaluation. In this work, we benchmark five common representations, Event Frame (EF), Voxel Grid (VG), Pseudo-Frames (PS), Image of Warped Events (IWE), and Event Spike Tensor (EST), across two datasets (VisEvent and LaSOT), using both pure event and hybrid RGB-event trackers. We find that representation choice significantly impacts performance, with EST and IWE consistently outperforming others, while EF and PS show poor robustness under distribution shifts. To address representation variability, we propose the Gradient-Unified Shared Embedding Module (GUSEM), a dual-pathway architecture that fuses heterogeneous event inputs into a shared, edge-aware embedding space. GUSEM leverages spatial gradients for structural consistency and low-rank reconstruction for modality-specific semantics. Extensive experiments show that GUSEM improves accuracy and generalization across trackers, representations, and training regimes, establishing it as a robust, representation-agnostic solution for event-based tracking.

1. Introduction

Visual Object Tracking (VOT) is a fundamental problem in computer vision that involves learning an appearance model for an arbitrary target object given only its initial state [17]. Its importance spans wide applications such as autonomous driving [4], and surveillance [1], to name a few. In recent years, numerous conventional RGB trackers have achieved significant performance improvements with the introduction of the Transformers [24] architecture to VOT [5, 6, 30, 31]. Additionally, the advent of large-scale VOT datasets such as LaSOT [7], LaSOT_{ext} [8], GOT-10k [16], and TrackingNet [22] has significantly facilitated the end-to-end training of these trackers. Despite these advance-

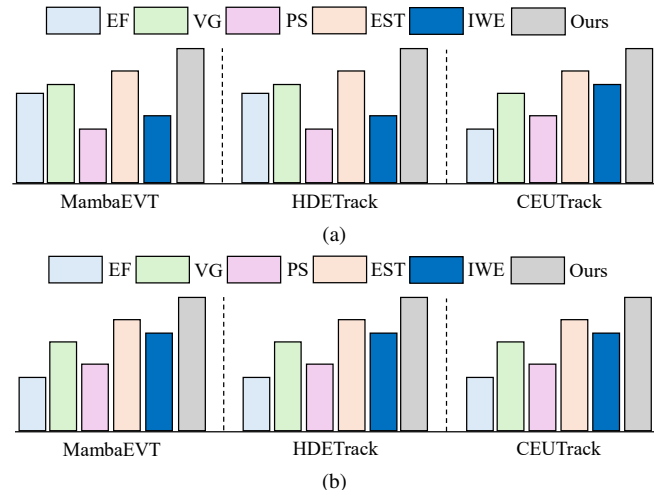


Figure 1. Success rate comparison of state-of-the-art (SOTA) trackers across different event image representations on the (a) VisEvent and (b) LaSOT datasets. “Ours” indicates each tracker equipped with the proposed Gradient-Unified Shared Embedding Module (GUSEM).

ments, contemporary trackers encounter persistent challenges, including Fast Motion (FM), Motion Blur (MB), and Illumination Variations (IV) [20], primarily due to their reliance on RGB cameras.

To overcome these challenges, researchers have explored enhancing input data effectiveness in VOT by introducing biologically inspired event cameras, such as Dynamic Vision Sensors (DVS) [3, 23, 25, 26, 28]. Unlike conventional RGB cameras that capture frames at fixed intervals, event cameras asynchronously capture per-pixel intensity changes and output a stream of events, making them immune to MB [11]. Each event encodes information including the timestamp, pixel location, and the polarity of the intensity change. As a result, event sensors offer ultra-high temporal resolution (μ s-level), a significantly higher dynamic range (140dB compared to 60dB in standard cameras) which enables them to work efficiently in poor IV. Additionally, DVS consumes far less energy and bandwidth while requiring minimal computational resources. These advantages, low latency, efficient resource utilization, and adaptability to diverse environments, make the event cameras well-suited for target tracking in challenging scenarios [3, 11, 15, 25, 27, 28, 32].

Despite their benefits, event cameras do not capture slow motion or static objects and lack fine-grained texture information which is also very important for high-performance tracking [11]. Therefore, the integration of visible-light cameras with DVS has emerged as a promising approach for enhancing reliability in VOT and utilizing the benefits of both modalities. Several studies [3, 23, 25, 26, 28] have explored this hybrid setting; however, many have arbitrarily chosen an event image representation without systematically evaluating its suitability for VOT tasks. Recent works [2, 3] have demonstrated that a well-designed event image representation can outperform the conventional approach of stacking events within fixed time intervals to generate event frames. Alternative representations such as event point clouds or Voxel Grids (VGs) may provide a more informative spatiotemporal encoding for improved VOT performance.

Exploring the event camera representation was already tackled by Zubic et al. [35], Gehric et al. [12] and Jiao et al. [19]. In those works event camera representation were investigated for object classification, object recognition, optical flow and SLAM respectively. However, to the best of our knowledge, this study has not been done for tracking applications. In addition, specifically for fusing the event data modality with RGB camera, an ideal representation that would enrich the tracker with features is essential.

In this work, we aim to address the gap in evaluating event image representations for VOT by conducting a comprehensive assessment of two pure event-based trackers, HDETrack [28], and MambaEVT [27], and one hybrid visible-event tracker, CEUTrack [23]. Our evaluation covers five event image representations: 1) Events Frame (EF), 2) Voxel Grid (VG), 3) Pseudo-Frames (PS), 4) Contrast Maximization, and 5) learnable representation, Event Spike Tensor (EST) [12], tested on two benchmark datasets: the real visible-event dataset VisEvent [25] and the synthetically generated visible-event dataset LaSOT [7], converted using V2E [14].

Beyond this evaluation, inspired by [29], we propose the Gradient-Unified Shared Embedding Module (GUSEM) to tackle representation-level heterogeneity in event-based tracking. GUSEM combines explicit gradient cues with implicit low-rank reconstruction to create a shared, edge-aware embedding space. It builds on the observation that edge structures, captured through spatial gradients, are the most stable features across diverse event image representations. Instead of rigid unification, GUSEM decomposes each representation into low-rank components via modality-specific multilayer perceptrons (MLPs). These are then fused and guided by the explicit gradient features to form a compact, modality-agnostic embedding. This dual-pathway design preserves both shared contours and representation-specific nuances, promoting robust general-

ization. This addition indicates an improvement in performance across all representations used as highlighted in Figs. 1(a) and (b) across the datasets VisEvent [25] and LaSOT [7], respectively.

2. Related Work

This work provides guidelines for choosing the most suitable event representation for VOT, focusing on methods that either rely solely on event cameras or fuse event and visible data, while excluding purely visible-based approaches.

2.1. Event Camera based Tracking

Event-based VOT has gained traction with the emergence of benchmark datasets tailored to raw event streams [23, 25, 26, 28]. Several approaches leverage different event representations and tracking paradigms: Mitrokhin et al. [21] model event geometry for motion compensation, while Jiang et al. [18] use event count images for robust detection and tracking. Fu et al. [10] propose DANet with a transformer-based Siamese framework, and Wang et al. [15] introduce MambaEVT using structured event frames. While event cameras excel in high-speed and HDR scenarios, they struggle with slow motion and texture-rich regions. Hybrid systems that combine event and RGB inputs have thus emerged as a promising solution for robust VOT.

2.2. Tracking by Combining Visible and Event Camera

Integrating visible and event sensors in VOT offers complementary strengths: RGB frames are valuable for static or slow-moving scenes, while event data excels under FM, MB, and IV [11, 23, 25, 26, 28]. Several datasets and frameworks have been introduced to support this hybrid paradigm.

Wang et al. [25] proposed the VisEvent dataset and a cross-modality Transformer for visible-event fusion, using event images. Tang et al. [23] introduced the COESOT dataset and a ViT-based tracker that integrates RGB with event voxels. Expanding on this, EventVOT [28] provides high-resolution visible-event pairs and a distillation-based tracker relying solely on event signals. FELT [26] targets long-term tracking with an associative memory Transformer using voxel-based events. Huang et al. [15] proposed Mamba-FETTrack, a State Space Model (SSM)-based framework combining RGB and event data via event images. Alansari et al. [3] presented ELTrack, a multi-modal tracker that uses pseudo-frames to fuse RGB, event, and language inputs efficiently.

Although prior work acknowledges that event representation impacts performance [2], no study has systematically compared different formats in a VOT context. Motivated by this gap, we evaluate five major event representations and

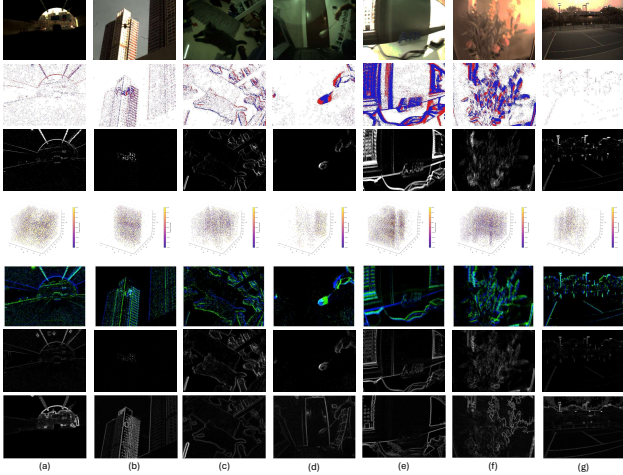


Figure 2. Sample frames from the VisEvent dataset illustrating various event image representations. Each column depicts a different scene, each row corresponds to a specific representation: (1) the original RGB image, (2) Event Frame (EF), (3) Pseudo-Frames (PS), (4) Voxel Grid (VG), (5) Event Spike Tensor (EST), (6) Image of Warped Events (IWE), and (7) our proposed Gradient-Unified Shared Embedding Module (GUSEM) output. The scenes include: (a) a vehicle in a tunnel, (b) a drone in front of a tall building, (c) a cat lying on a surface, (d) an indoor scene with a partially visible person, (e) a logo or letter next to a computer monitor with visible cables, (f) a scene with flowers, and (g) an outdoor tennis court or sports area.

their influence on tracking performance to establish clearer guidelines for future research and deployment.

3. Method

To evaluate the effectiveness of different event image representations (see Fig. 2) for VOT, we design a standardized framework in which each representation is used as input to a shared baseline tracker. Then, we introduce a lightweight, representation-agnostic modification to any existing tracker, termed the GUSEM shown in Fig. 3, which enhances performance across all input types.

3.1. Input Representation

Event cameras consist of independent pixels that asynchronously monitor changes in the logarithmic intensity of light, denoted by $L(x, y, t)$, at each spatial location (x, y) . An event is triggered when the change in logarithmic intensity exceeds a predefined threshold C , formulated as:

$$\Delta L(x, y, t) = \log(L(x, y, t)) - \log(L(x, y, t - \Delta t)) \geq pC \quad (1)$$

where $p \in \{-1, 1\}$ is the polarity of the change in brightness which indicates whether the intensity increased (+1) or decreased (-1), and Δt represents the time elapsed since the last recorded event at that same pixel location (x, y) . In

a given time interval $\Delta\tau$, the event camera generates a sequence of asynchronous events forming an unordered set:

$$\mathcal{E} = \{e_i\}_{i=1}^N = \{(x_i, y_i, t_i, p_i)\}_{i=1}^N. \quad (2)$$

Each event e_i encodes the spatiotemporal coordinates (x_i, y_i, t_i) and polarity $p_i \in \{-1, 1\}$, representing a change in brightness at pixel (x_i, y_i) and time t_i . These events are sparse, high-temporal-resolution, and inherently different from conventional frame-based data, necessitating a transformation into structured representations for use in learning-based VOT systems.

3.2. Event Image Representations

To make raw event data compatible with deep learning-based VOT methods, which require dense, grid-based inputs, the asynchronous event stream must first be transformed into structured image-like representations. Given an unordered set of events Eq. (2) collected within a fixed temporal window $\Delta\tau$ or a fixed number of events, we construct dense tensors $I \in \mathbb{R}^{H \times W \times C}$, where H and W denote the spatial resolution of the sensor, and C corresponds to the number of channels specific to each representation. These representations serve as the input to standard convolutional or transformer-based trackers.

Each representation used in this study (Fig. 2) is designed to visualize the raw event stream while either preserving or discarding temporal information, depending on its formulation. Sample visualizations of these representations are provided in Fig. 2. For consistency, all event data in our experiments are discretized using a fixed temporal window aligned with the timestamp of the corresponding RGB frame. In this work, we evaluate several commonly used event image representations:

Event Frame: An *Event Frame (EF)* is a spatial-only representation of events that occur within a user-defined time window Δt (see Fig. 2, second row). Events detected during this interval are mapped onto a frame with the same resolution as the camera sensor. This representation disregards the temporal aspect of the data, meaning that if multiple events occur at the same spatial location within the same time window, only the most recent event will be retained, effectively overwriting previous ones. To preserve polarity information, a two-channel representation is used, where events with positive and negative polarity are accumulated separately. This ensures that both positive and negative events contribute distinctly to the final representation.

Pseudo-Frames: A *Pseudo-Frames (PS)* representation is a spatial-only representation of events that occur within a user-defined time window Δt (see Fig. 2, third row).

begins with generating a dense event stream (e.g., via the *Video-to-Events (V2E)* module [14]), followed by noise

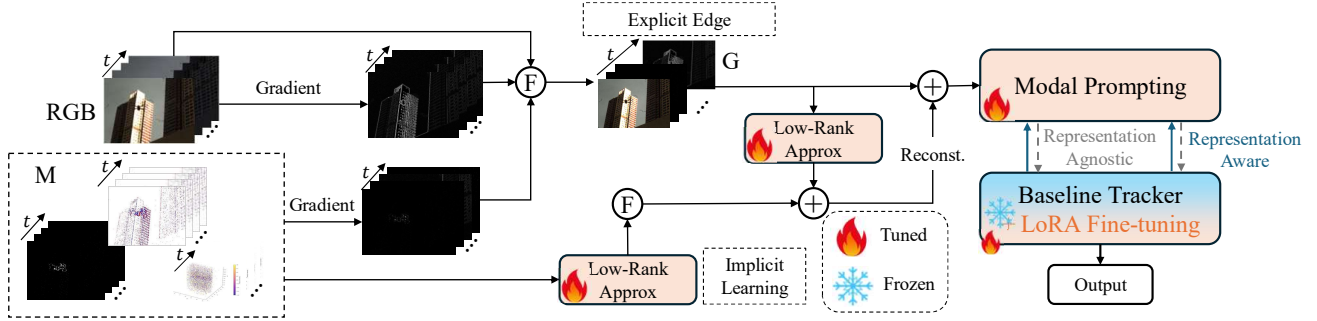


Figure 3. Overview of our proposed GUSEM-based tracking framework. It consists of a module for unifying event representations, a modal prompting block for adaptive token refinement, and a LoRA-tuned RGB tracker. The unified design enables robust tracking across diverse event formats with a single model.

filtering and accumulation. To enhance robustness, the *Yang filter* [9] eliminates isolated or spurious events by checking local spatiotemporal consistency:

$$E'(x, y, t) = \begin{cases} E(x, y, t), & \text{if } \sum_{(x', y', t') \in \mathcal{N}} \mathbb{I}(E(x', y', t') \neq 0) \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Subsequently, the filtered events are integrated over a fixed time window to form the final PS:

$$P(x, y) = \sum_{t=t_0}^{t_0+T} E'(x, y, t), \quad (4)$$

where $P(x, y)$ represents the accumulated event count at each pixel over the interval T .

Voxel Grid: A *Voxel Grid (VG)* is a three-dimensional (3D) spatiotemporal representation of events, where each voxel corresponds to a specific pixel location and a discrete time interval (see Fig. 2, fourth row). A VG divides the event stream into B temporal bins (or slices) over a time window Δt . Each event is assigned to a voxel based on its spatial coordinates (x, y) and its relative timestamp within the window. To mitigate quantization artifacts and improve the smoothness of the representation, bilinear interpolation is applied. This allows each event to contribute proportionally to neighboring spatial pixels and temporal bins, with weights determined by its distance from the voxel centers [33]. Such soft-assignment preserves finer motion patterns and ensures continuity across voxel boundaries [34].

The VG is computed following the formulation of Zhu et al. [34]. First, the temporal position of each event is normalized to the bin index range:

$$t_i^* = (B - 1) \frac{(t_i - t_1)}{(t_N - t_1)} \quad (5)$$

Then, the VG is constructed by accumulating event contri-

butions using a bilinear kernel:

$$V(x, y, t) = \sum_i p_i k_b(x - x_i) k_b(y - y_i) k_b(t - t_i^*) \quad (6)$$

where $p_i \in \{-1, +1\}$ is the event polarity, and the bilinear kernel $k_b(\cdot)$ is defined as $k_b(a) = \max(0, 1 - |a|)$. This results in a dense 3D tensor $V \in \mathbb{R}^{B \times H \times W}$, which retains temporal ordering while remaining compatible with standard convolutional architectures.

Event Spike Tensor: The *Event Spike Tensor (EST)* is a four-dimensional spatiotemporal representation introduced by Gehrig et al. [12] (see Fig. 2, fifth row), designed to preserve the fine temporal resolution of event data while maintaining dense spatial structure. Unlike discretized representations such as VG, which rely on fixed temporal binning, EST encodes time continuously using differentiable temporal basis functions. These can be either fixed (e.g., raised cosines) or learned via a MLP that maps event timestamps and coordinates into a high-dimensional space. Each event's response is then projected onto a structured grid and accumulated across channels, forming a smooth, multi-channel tensor that avoids the quantization artifacts typical of hard-binning approaches.

This soft temporal encoding produces smooth temporal profiles at each pixel, enabling the capture of high-frequency motion patterns and subtle temporal variations. Because the EST is fully differentiable and can be trained end-to-end, it allows the representation itself to adapt to the requirements of downstream tasks such as tracking. As a result, EST offers a flexible, information-preserving framework that effectively bridges raw event streams and deep neural architectures.

Image of Warped Events: The Contrast Maximization (CM) is a general-purpose framework for extracting motion or structure from event data by exploiting the principle that events are triggered at locations of high temporal contrast.

Given a batch of events $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^N$, CM seeks to align them along a hypothesized motion model parameterized by θ . Each event is warped to a common reference time t_{ref} according to this motion model, resulting in warped coordinates:

$$\mathbf{x}'_i(\theta) = \mathcal{W}(\mathbf{x}_i, t_i; \theta), \quad (7)$$

where $\mathcal{W}(\cdot)$ denotes the warping function that compensates for motion between t_i and t_{ref} .

The warped events are accumulated into a 2D image called the *Image of Warped Events (IWE)* (see Fig. 2, sixth row):

$$H(\mathbf{x}; \theta) = \sum_{i=1}^N k(\mathbf{x} - \mathbf{x}'_i(\theta)), \quad (8)$$

where $k(\cdot)$ is a spatial kernel (e.g., a box or Gaussian kernel) used to smooth the event contributions onto the image plane.

The goal is to find the motion parameters θ that maximize a contrast-based objective function over the IWE. A commonly used contrast function is the variance of the pixel intensities:

$$f(\theta) = \text{Var}(H(\mathbf{x}; \theta)). \quad (9)$$

Intuitively, when the correct motion parameters are used, the warped events align coherently, resulting in a sharper, high-contrast IWE. The optimization $\theta^* = \arg \max_{\theta} f(\theta)$, yields motion estimates such as optical flow, object trajectory, or camera motion.

3.3. GUSEM

Event image representations such as EF, VG, PS, EST, and IWE vary in their encoding of spatial and temporal dynamics. Despite this heterogeneity, a consistent structural prior emerges across formats: object contours and boundary transitions. To unify these modalities into a robust and generalizable representation space, we propose the GUSEM, inspired by the modality alignment strategy introduced in [29]. GUSEM combines explicit gradient-based structural supervision with implicit low-rank factorization to capture shared semantics while preserving modality-specific subtleties (see Fig. 2, seventh row).

Explicit Structural Unification: While each representation encodes event data differently, e.g., temporal binning in EF, voxel stacking in VG, filtering in PS, spike-driven activation in EST, or motion-compensated warping in IWE, they all retain edge-related structure due to the inherent nature of events triggering on brightness change. This results in a shared structural prior: edges and motion boundaries consistently emerge across formats.

GUSEM exploits this inherent property by computing spatial gradient maps for each representation using fixed

horizontal and vertical finite differences. This yields a consistent, modality-agnostic structural cue G , which serves as an explicit edge-aware signal aligned with high-frequency features such as contours and motion boundaries. To retain this structure across network depth, we concatenate the gradient maps G with the original input features, forming a fused representation that injects edge-awareness early and preserves it through deeper layers. This design ensures that edge fidelity, a key asset in event-based vision, is maintained and aligned across representations, laying the groundwork for subsequent shared embedding learning.

Implicit Low-Rank Structural Embedding: While gradient-based features capture shared structural priors (e.g., edges, contours), they lack the capacity to model representation-specific variations inherent in different event formats. To complement this, GUSEM integrates an implicit low-rank learning path that captures nuanced differences while preserving alignment with structural cues. Specifically, input features M from multiple event image representations are decomposed into subsets $\{M^{\text{EF}}, M^{\text{VG}}, M^{\text{PS}}, M^{\text{IWE}}\}$. Each is processed through a modality-specific projection function σ_x to produce a compressed low-rank matrix $M_k^x = \sigma_x(M^x)$, where $k \ll c$ and $\sigma_x : \mathbb{R}^c \rightarrow \mathbb{R}^k$ is implemented as a lightweight MLP. Simultaneously, the explicit gradient map G is projected into the same low-rank space $G_k = \sigma_g(G)$ to act as a structural prior.

We then construct a joint latent embedding by concatenating the per-representation low-rank matrices and integrating the gradient prior:

$$M_k = \varphi_{R_1}([M_k^{\text{EF}}, M_k^{\text{VG}}, M_k^{\text{PS}}, M_k^{\text{EST}}, M_k^{\text{IWE}}]) + \varphi_{R_2}(G_k), \quad (10)$$

where φ_{R_1} and φ_{R_2} are learnable fusion MLPs. Finally, the unified low-rank representation M_k is projected back to the original feature space and combined with the gradient feature to form the final shared embedding:

$$F = \Phi_R(M_k) + G, \quad (11)$$

where Φ_R is a lightweight reconstruction MLP, and the residual connection to G reinforces structural consistency. This dual pathway, explicit structural supervision and implicit low-rank learning, allows GUSEM to construct a robust, semantically-aligned embedding space that preserves both global structure and format-specific distinctions across event representations.

Outer Representation Prompting: Event representations often encode complementary cues, some emphasize fine-grained temporal precision (e.g., EST, VG), while others prioritize spatial consistency (e.g., EF, PS, IWE). A pre-trained RGB tracker or ViT operating on a single representation (e.g., EF) may struggle to generalize across corner

cases. To address this, we introduce a representation-aware prompting mechanism that adaptively enhances input tokens using cross-representation cues.

Inspired by UnTrack [29], prompting and LoRA-based adaptation techniques, we implement a shrinkage token fusion strategy over input tokens I . Each token is dynamically scored by a function $s(\cdot)$, which segments the token space into: (1) positive tokens (m_p) which are confident and reliable, (2) uncertain tokens (m_u) which are ambiguous or noisy, and (3) negative tokens (m_n) which are low-confidence or degraded.

The goal is to inject useful structure from the shared embedding F into the base tokens, while preserving confident ones. Negative tokens are replaced by corresponding tokens from F , and positive tokens are retained from I . These are linearly projected to a low-rank space:

$$I_{l_1} = \sigma_c(m_n \cdot F + m_p \cdot I), \quad (12)$$

where σ_c is a projection MLP to the low-rank space.

Next, we focus on uncertain tokens. These are softly fused with their counterparts in F , with the aim of suppressing noise and enhancing signal:

$$I_{l_2} = \sigma_n(m_u \cdot F + m_u \cdot I), \quad (13)$$

where σ_n is another low-rank projection function. The two branches are then combined:

$$I_l = \varphi_P([I_{l_1}, I_{l_2}]), \quad (14)$$

producing a unified low-rank embedding of the input tokens. A similar pipeline is applied to F to obtain its low-rank representation F_l . Finally, we perform cross-representation fusion:

$$O = \Phi_P(I_l + F_l), \quad (15)$$

where Φ_P is an MLP that projects back to the original feature space, producing the prompted, representation-aware output.

This mechanism serves as a token-level filter and enhancer: (1) negative tokens are overwritten with more trustworthy signals, (2) uncertain tokens are denoised through fusion, and (3) positive tokens are left untouched. Because all fusion operations occur in a compressed low-rank space, the prompting adds minimal overhead while enabling dynamic and selective enhancement of event representations.

Inner Representation Prompting: While outer prompting enriches input tokens using cross-representation cues, deeper adaptation within the model is often necessary to fully align the network with the diverse statistical patterns of event-based representations. However, fine-tuning the entire network is computationally expensive and prone to

Table 1. Per-representation fine-tuning performance comparison of various event image representations on the VisEvent and LaSOT datasets.

Method	Rep.	VisEvent				LaSOT				
		S	NP	P	Δ	S	NP	P	Δ	
Event Only	MambaEVT	VG	36.2	39.7	50.8	-2.3	25.6	26.3	24.7	-2.9
		EF	35.9	39.4	50.2	-2.6	21.6	22.0	20.0	-6.9
		PS	32.9	36.7	46.3	-5.6	22.9	23.4	21.2	-5.6
		EST	36.6	40.2	51.6	-1.9	26.7	28.2	25.2	-1.8
		IWE	33.8	37.5	47.2	-4.7	26.1	27.7	24.8	-2.4
	Ours	38.5	42.7	53.4	-	28.5	30.4	28.6	-	
	HDETrack	VG	37.3	44.5	54.6	-2.2	26.1	28.6	25.0	-3.9
		EF	36.3	43.4	53.8	-3.2	23.2	24.0	20.1	-6.8
		PS	33.8	41.2	51.6	-5.7	23.9	24.6	20.4	-6.1
		EST	37.8	44.9	54.9	-1.7	27.6	29.5	26.4	-2.4
IWE		35.7	42.8	52.5	-3.8	27.1	29.2	26.1	-2.9	
Ours	39.5	46.4	56.6	-	30.0	32.3	30.4	-		
RGB & Event	CEUTrack	VG	53.1	73.8	69.1	-3.6	39.5	45.8	37.1	-4.6
		EF	51.5	69.7	67.5	-5.2	37.8	44.6	35.3	-6.3
		PS	52.8	71.6	68.7	-3.9	38.5	44.9	36.6	-5.6
		EST	54.9	74.5	70.7	-1.8	42.3	47.7	39.4	-1.8
		IWE	54.4	74.1	70.5	-2.3	41.6	46.7	38.4	-2.5
		Ours	56.7	76.3	72.4	-	44.1	49.3	41.4	-

overfitting, especially when working with sparse, high-resolution event data. To balance adaptability and efficiency, we incorporate Low-Rank Adaptation (LoRA) [13] into the model’s core transformer layers.

Specifically, for each attention module with a frozen weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we introduce two trainable matrices: low rank projection $A \in \mathbb{R}^{r \times k}$, and low-rank expansion $B \in \mathbb{R}^{d \times r}$, where $r \ll \min(d, k)$ ensures minimal parameter overhead. The original attention operation $h = W_0x$ is augmented as $h = W_0x + BAx$. The LoRA term BAx injects a small, trainable adaptation into the frozen transformer without modifying its pre-trained weights. This allows the model to internally adjust to representation-specific patterns, such as the sparse structure of EST, the smooth interpolation of PS, or the motion-aligned textures of IWE, without disrupting the core RGB tracking functionality.

We optimize the network end-to-end using the same loss objectives as our baseline tracker, enabling seamless integration with existing training pipelines. By confining trainable parameters to LoRA modules, prompting layers, and the shared embedding module, the entire system remains GPU-friendly, scalable, and robust to overfitting.

4. Experiments

4.1. Implementation Details

We evaluate tracking performance on two datasets: VisEvent for Short-Term (ST) event-based tracking and LaSOT (with synthetic events from V2E) for Long-Term (LT) scenarios. VisEvent includes 820 real event sequences with diverse motion and lighting challenges, while LaSOT offers 1,400 RGB videos converted to events for LT evaluation. We benchmark three baseline trackers (CEUTrack, MambaEVT, HDETrack) across five event representations, with

Table 2. Joint representation performance of trackers trained on all event image representations and tested separately on each.

Method	Rep.	VisEvent				LaSOT				
		S	NP	P	Δ	S	NP	P	Δ	
Event Only	MambaEVT	VG	30.8	34.3	45.4	-6	23.3	23.9	22.1	-3.8
		EF	29.5	33.7	44.1	-7.3	19.8	20.0	18.5	-7.3
		PS	29.2	33.4	43.9	-7.6	20.5	20.8	18.8	-6.6
		EST	31.9	35.5	46.9	-4.9	24.1	25.9	22.8	-3
		IWE	30.3	34.1	45.2	-6.5	23.8	25.6	22.4	-3.3
	Ours	36.8	41.5	52.6	-	27.1	29.5	27.8	-	
	HDETrack	VG	34.6	41.7	52.0	-3.4	24.3	26.2	23.4	-4.5
		EF	31.2	38.5	48.5	-6.8	21.4	22.1	18.1	-7.4
		PS	30.8	38.1	48.7	-7.2	21.9	22.5	18.5	-6.9
		EST	35.1	42.3	52.1	-2.9	25.1	26.8	24.1	-3.7
IWE		33.3	40.6	50.2	-4.7	24.7	26.6	23.8	-4.1	
Ours	38.0	45.1	55.2	-	28.8	31.2	29.4	-		
RGB & Event	CEUTrack	VG	51.3	71.0	67.6	-4	39.1	44.6	36.4	-3.7
		EF	49.1	68.4	65.9	-6.2	36.7	43.8	34.7	-6.1
		PS	50.2	69.5	66.8	-5.1	37.3	44.1	35.0	-5.5
		EST	52.0	71.6	68.2	-3.3	40.4	45.5	37.2	-2.4
		IWE	51.6	71.2	67.8	-3.7	39.4	44.9	36.6	-3.4
		Ours	55.3	75.0	71.1	-	42.8	47.9	40.2	-

and without our proposed GUSEM.

All experiments are conducted on a single NVIDIA GeForce RTX 3080 GPU. For each setup, fine-tuning is performed for 10 epochs using the tracker’s default training configuration and initialized from its official pretrained weights. Tracking performance is evaluated using standard metrics: Success (S), Precision (P), and Normalized Precision (NP). To comprehensively evaluate the generalization and representation-specific performance of trackers, we design three experimental protocols:

Per-representation Fine-tuning Setting: Each tracker is fine-tuned individually on a specific event image representation before being evaluated on the same representation.

Joint Representation Training Setting: Each tracker is trained using a mixed dataset that includes all considered representations. This setup evaluates the ability of the tracker to generalize across heterogeneous inputs using a unified model.

Cross-Representation Evaluation Setting: A tracker trained on one representation is directly tested on a different representation without further fine-tuning. This tests the transferability and robustness of learned features across representations.

4.2. Results

To analyze the impact of training strategies on representation-specific and cross-representation performance, we evaluate our model under three experimental settings described in Section 4. Below, we report and interpret the results obtained for each. For each tracker, the top three results are highlighted in **red** (best), **green** (second), and **blue** (third). *Ours* refers to the corresponding tracker equipped with our proposed GUSEM. *Rep.* denotes the event image representation used as input, and Δ indicates the difference in S score between each method and its GUSEM-enhanced counterpart.

Per-Representation Fine-Tuning: Table 1 shows how different trackers perform when fine-tuned and evaluated on individual event representations across VisEvent and LaSOT. This setup reveals how well each format supports tracking when used in isolation.

Among baseline methods, EST consistently yields the strongest results, particularly on VisEvent, due to its learnable, continuous-time encoding that captures fine temporal structure. VG also performs well, especially with MambaEVT, owing to effective discretization and interpolation. IWE trails closely behind, leveraging motion-aligned accumulation for robust structure preservation. In contrast, EF and PS underperform, EF due to coarse quantization, and PS due to information loss from filtering, particularly on LaSOT, which contains more variation and longer sequences.

The Δ column quantifies improvements from replacing each baseline representation with our GUSEM. On MambaEVT, GUSEM outperforms EST by +1.9, +2.5, +1.8 (VisEvent) and +1.8, +2.2, +3.4 (LaSOT). HDETrack sees similar gains: +1.7, +1.5, +1.7 on VisEvent and up to +4.0 on LaSOT. CEUTrack achieves the highest scores overall, with gains up to +2.6 over EST and IWE. These results highlight GUSEM’s ability to unify structurally diverse inputs into a stable, edge-aware embedding that generalizes across scenes and sequence lengths. The consistent performance boost across all trackers validates our dual-path design: explicit gradient supervision for structural consistency and low-rank fusion for semantic alignment.

Joint Representation Training: To evaluate generalization under heterogeneous inputs, we adopt a joint training setup using all five event representations (EF, VG, PS, EST, IWE). As shown in Table 2, GUSEM-enhanced models consistently outperform their vanilla counterparts across all trackers and datasets.

Among standard formats, EST and IWE remain the most effective. EST achieves the highest scores overall due to its learnable, continuous encoding, while IWE follows closely, leveraging motion-aligned accumulation. VG and PS generalize less effectively, particularly on LaSOT, likely due to quantization and temporal sparsity. EF performs worst, limited by its weak temporal fidelity and structural degradation.

The Δ column shows GUSEM’s margin over baselines. On MambaEVT, GUSEM improves over EST by +4.9 (S), +6.0 (NP), +5.7 (P) on VisEvent, and +3.0 to +4.0 on LaSOT. HDETrack sees +2.9-3.0 gains on VisEvent and up to +5.3 on LaSOT. CEUTrack also benefits, with gains up to +2.4 over EST and IWE. These results demonstrate that while joint training helps, conventional trackers are still limited by their input formats. In contrast, GUSEM unifies diverse representations into a shared embedding that is structurally aligned and semantically rich. Its explicit gradient cues and low-rank fusion jointly enable strong gener-

Table 3. Cross-representation tracking performance on VisEvent dataset.

Method	Train \ Test	EF	VG	PS	EST	IWE	Avg.
MambaEVT	EF	35.9	35.1	32.3	31.2	33.4	33.6
	VG	35.5	36.2	33.8	35.6	33.4	34.9
	PS	30.0	30.7	32.9	31.1	31.8	31.3
	EST	33.9	34.5	33.3	36.6	33.6	34.4
	IWE	31.4	31.7	32.5	32.1	33.8	32.3
	Ours	38.5	38.5	36.8	37.1	36.9	37.6
HDETrack	EF	36.3	35.9	33.8	35.4	34.2	35.1
	VG	36.6	37.3	35.1	36.1	35.6	36.1
	PS	30.3	31.0	33.8	31.5	32.7	31.9
	EST	35.7	36.6	35.1	37.8	35.3	36.1
	IWE	32.1	32.3	34.9	34.2	35.7	33.8
	Ours	39.5	39.5	37.1	38.2	37.7	38.4
CEUTrack	EF	51.5	51.1	50.6	51.0	50.9	51.0
	VG	52.7	53.1	52.5	52.9	52.6	52.8
	PS	51.3	51.6	52.8	51.9	52.1	51.9
	EST	53.3	53.9	53.4	54.9	53.4	53.8
	IWE	53.2	53.4	54.1	53.7	54.4	53.8
	Ours	56.7	56.7	56.3	56.4	56.6	56.5

alization across conditions.

Cross-Representation Evaluation: To evaluate robustness under representation shifts, we conduct cross-representation tests where models are trained on one event format and tested on others. This setup mimics real-world deployment, where the input encoding may vary or be unknown. As shown in Tables 3 and 4, several consistent trends emerge.

Trackers perform best on their training format (diagonal entries), but performance often drops on mismatched formats, revealing overfitting to representation-specific traits. Among standard formats, EST and IWE exhibit the strongest generalization, yielding higher off-diagonal scores due to their stable structural encoding. In contrast, EF and PS perform poorly across representations, especially on LaSOT, highlighting their weak temporal modeling and limited adaptability. LaSOT exposes these weaknesses more clearly: EF- and PS-trained models drop by 5-7 points across test formats, while VG and EST degrade more moderately. Still, no conventional representation achieves consistent cross-format performance.

GUSEM overcomes these limitations across all trackers. On VisEvent, MambaEVT with GUSEM averages 37.6, improving over EST by +2.7; HDETrack reaches 38.4 vs. 36.1 (VG/EST). Even CEUTrack, with strong baselines, gains +2.7 with GUSEM. On LaSOT, GUSEM improves MambaEVT from 25.2 (IWE) to 27.8, HDETrack from 25.7 to 29.3, and CEUTrack from 40.1 to 43.5. These results confirm GUSEM’s ability to unify heterogeneous inputs into a structurally stable, semantically rich embedding. Its explicit gradient path captures invariant contours, while low-rank fusion models transferable representation-specific cues.

Table 4. Cross-representation tracking performance on LaSOT dataset.

Method	Train \ Test	EF	VG	PS	EST	IWE	Avg.
MambaEVT	EF	21.6	20.6	18.5	19.7	16.3	19.3
	VG	24.2	25.6	20.4	23.8	22.4	23.3
	PS	19.5	20.4	22.9	21.3	21.7	21.2
	EST	24.7	25.6	24.5	26.7	24.4	22.5
	IWE	23.1	23.8	24.8	25.1	26.1	25.2
	Ours	28.5	28.5	27.3	27.1	27.6	27.8
HDETrack	EF	23.2	21.4	20.1	21.5	18.7	21.0
	VG	25.7	26.1	23.4	25.1	22.6	24.6
	PS	21.3	21.5	23.9	20.4	22.6	21.9
	EST	24.1	25.8	24.5	27.6	24.7	25.3
	IWE	24.6	25.1	25.7	25.9	27.1	25.7
	Ours	30.0	30.0	28.8	28.5	29.1	29.3
CEUTrack	EF	37.8	37.1	35.2	36.6	35.5	36.4
	VG	38.7	39.5	37.4	38.9	37.7	38.4
	PS	35.6	35.9	38.5	37.1	37.8	37.0
	EST	39.7	41.2	38.4	42.3	38.9	40.1
	IWE	38.1	38.4	39.6	39.9	41.6	39.5
	Ours	44.1	44.1	43.2	42.7	43.3	43.5

5. Conclusion

This work presents a systematic analysis of event image representations in VOT. Our evaluation across five representations, EF, VG, PS, IWE, and EST, demonstrates that performance varies widely depending on the chosen format. No single representation proves universally effective, which limits the reliability of existing event-based tracking methods. To address this challenge, we introduce GUSEM designed to unify diverse event formats into a consistent representation space. GUSEM integrates two complementary pathways: explicit spatial gradients that capture modality-invariant structural cues, and low-rank reconstruction that preserves format-specific semantics. Experimental results across multiple trackers and datasets show that GUSEM consistently improves tracking accuracy under fine-tuned, joint training, and cross-representation conditions. By decoupling performance from the specific event encoding, GUSEM offers a robust and flexible solution for real-world deployment where input representations may vary.

6. Acknowledgment

This work is supported by Sandoq Al Watan, United Arab Emirates under Grant SWARD-S22-015. This work is also supported by the Advanced Research and Innovation Center (ARIC), which is jointly funded by the Aerospace Holding Company LLC, a wholly-owned subsidiary of Mubadala Investment Company PJSC, United Arab Emirates and Khalifa University of Science and Technology, United Arab Emirates.

References

- [1] Mohamad Alansari, Oussama Abdul Hay, Sara Alansari, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. Drone-person tracking in uniform appearance crowd: A new dataset. *Scientific Data*, 11(1):15, 2024. 1
- [2] Mohamad Alansari, Hamad AlRemeithi, Sara Alansari, Naoufel Werghi, and Sajid Javed. Performance analysis of synthetic events via visual object trackers. In *Intelligent Computing*, pages 364–384, Cham, 2024. Springer Nature Switzerland. 2
- [3] Mohamad Alansari, Khaled Alnuaimi, Sara Alansari, and Sajid Javed. Eltrack: Events-language description for visual object tracking. *IEEE Access*, 2025. 1, 2
- [4] Fei Chen, Xiaodong Wang, Yunxiang Zhao, Shaoh Lv, and Xin Niu. Visual object tracking: A survey. *Computer Vision and Image Understanding*, 222:103508, 2022. 1
- [5] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13608–13618, 2022. 1
- [6] Yutao Cui, Cheng Jiang, Gangshan Wu, and Limin Wang. Mixformer: End-to-end tracking with iterative mixed attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4129–4146, 2024. 1
- [7] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2
- [8] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129: 439–461, 2021. 1
- [9] Yang Feng, Hengyi Lv, Hailong Liu, Yisa Zhang, Yuyao Xiao, and Chengshan Han. Event density based denoising method for dynamic vision sensor. *Applied Sciences*, 10(6), 2020. 4
- [10] Yingkai Fu, Meng Li, Wenxi Liu, Yuanchen Wang, Jiqing Zhang, Baocai Yin, Xiaopeng Wei, and Xin Yang. Distractor-aware event-based tracking. *IEEE Transactions on Image Processing*, 32:6129–6141, 2023. 2
- [11] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. 1, 2
- [12] 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. 2, 4
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6
- [14] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 2, 3
- [15] Ju Huang, Shiao Wang, Shuai Wang, Zhe Wu, Xiao Wang, and Bo Jiang. Mamba-fetrack: Frame-event tracking via state space model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 3–18. Springer, 2024. 1, 2
- [16] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. 1
- [17] Sajid Javed, Martin Danelljan, Fahad Shahbaz Khan, Muhammad Haris Khan, Michael Felsberg, and Jiri Matas. Visual object tracking with discriminative filters and siamese networks: a survey and outlook. *IEEE TPAMI*, 45(5):6552–6574, 2022. 1
- [18] Rui Jiang, Xiaozheng Mou, Shunshun Shi, Yueyin Zhou, Qinyi Wang, Meng Dong, and Shoushun Chen. Object tracking on event cameras with offline-online learning. *CAAI Transactions on Intelligence Technology*, 5(3):165–171, 2020. 2
- [19] Jianhao Jiao, Huaiyang Huang, Liang Li, Zhijian He, Yilong Zhu, and Ming Liu. Comparing representations in tracking for event camera-based slam. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1369–1376, 2021. 2
- [20] Janani Kugarajeevan, Thanikasalam Kokul, Amirthalingam Ramanan, and Subha Fernando. Transformers in single object tracking: An experimental survey. *IEEE Access*, 11: 80297–80326, 2023. 1
- [21] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9, 2018. 2
- [22] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1
- [23] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*, 2022. 1, 2
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1
- [25] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Vi-sevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3): 1997–2010, 2023. 1, 2

- [26] Xiao Wang, Ju Huang, Shiao Wang, Chuanming Tang, Bo Jiang, Yonghong Tian, Jin Tang, and Bin Luo. Long-term frame-event visual tracking: Benchmark dataset and baseline. *arXiv preprint arXiv:2403.05839*, 2024. [1](#), [2](#)
- [27] Xiao Wang, Chao wang, Shiao Wang, Xixi Wang, Zhicheng Zhao, Lin Zhu, and Bo Jiang. Mambaevt: Event stream based visual object tracking using state space model. *CoRR*, abs/2408.10487, 2024. [1](#), [2](#)
- [28] Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19248–19257, 2024. [1](#), [2](#)
- [29] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19156–19166, 2024. [2](#), [5](#), [6](#)
- [30] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457, 2021. [1](#)
- [31] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022. [1](#)
- [32] Yajing Zheng, Zhaofei Yu, Song Wang, and Tiejun Huang. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. *IEEE Transactions on Image Processing*, 32:335–349, 2023. [1](#)
- [33] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 989–997, 2019. [4](#)
- [34] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019. [4](#)
- [35] Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12846–12856, 2023. [2](#)