Match-Any-Events: Zero-Shot Motion-Robust Feature Matching Across Wide Baselines for Event Cameras

Ruijun Zhang[1], Hang Su[2], Kostas Daniilidis[3,4], and Ziyun Wang[1]

[1] Johns Hopkins University, Baltimore MD 21218, USA
[2] ShanghaiTech University, Shanghai, China
[3] University of Pennsylvania, Philadelphia PA 19104, USA
[4] Archimedes, Athena RC, Greece

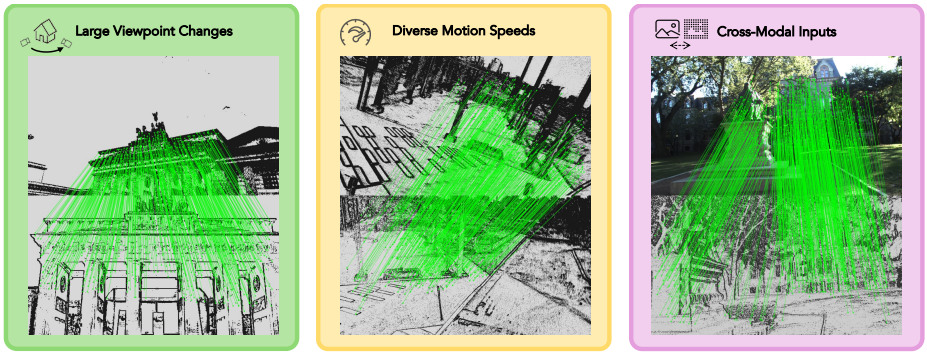

Fig. 1: Match-Any-Events can perform robust feature matching between two **arbitrary views** with different **motion profiles** and **across modalities**. No fine-tuning is needed for new datasets.

Abstract. Event cameras have recently shown promising capabilities in instantaneous motion estimation due to their robustness to low light and fast motions. However, computing wide-baseline correspondence between two arbitrary views remains a significant challenge, since event appearance changes substantially with motion, and learning-based approaches are constrained by both scalability and limited wide-baseline supervision. We therefore introduce the first event matching model that achieves cross-dataset wide-baseline correspondence in a **zero-shot manner**: a single model trained once is deployed on unseen datasets without any target-domain fine-tuning or adaptation. To enable this capability, we introduce a motion-robust and computationally efficient attention backbone that learns multi-timescale features from event streams, augmented with sparsity-aware event token selection, making large-scale training on diverse wide-baseline supervision computationally feasible. To provide the supervision needed for wide-baseline generalization, we develop a robust event motion synthesis framework to generate large-scale event-matching datasets with augmented viewpoints, modalities, and motions. Extensive experiments across multiple benchmarks show that our framework achieves a **37.7**% improvement over the previous best event feature matching methods. *Code and data are available at:* https://github.com/spikelab-jhu/Match-Any-Events.

1 Introduction

Biological vision systems exhibit remarkable abilities to establish correspondence across different spatial and temporal scales. At short spatiotemporal intervals, instantaneous motions of objects are captured by optical flow processed within motion-sensitive areas, such as Middle Temporal (MT) area in the brain [9, 26]. On the other hand, humans can easily establish correspondence between two distinct views with minimal overlaps, even in the absence of temporal continuity between two observations [17, 21, 45, 50]. This intrinsic understanding of correspondence enables self-localization, loop-closure in odometry, and recognition of structural and semantic similarities between objects. Recently, frame-based computer vision algorithms have made substantial progress for both instantaneous and wide-baseline matching, driven by large-scale datasets and scalable network architectures [41, 43, 54, 63, 66, 69, 71].

In contrast, biologically inspired event-based cameras have demonstrated **unbalanced** performance between **instantaneous** and **wide-baseline** correspondence. Event algorithms excel at estimating instantaneous motions due to their high temporal resolution and dynamic range, allowing them to handle extremely dynamic and event deformable objects [25, 74–77, 81, 85]. However, event-based methods have not shown generalizable performance in wide-baseline matching for several crucial factors. First, event data processing usually relies on a **handcrafted** encoding function that depends on hyperparameters such as time interval and number of events. These parameters change the appearance of events based on motions, which prevents networks trained on certain motion speeds from generalizing to different motions. Second, we do not have sufficient scale for the current event-based matching networks and datasets. On the data side, existing event datasets provide narrow-baseline optical flow ground truth, annotated with SLAM-derived poses and point maps. In our experiments, networks trained with only such data cannot generalize to large viewpoint changes. The need for diverse motion augmentations further amplifies the data problem. On the model side, dense spatiotemporal attentions in event transformers increase the attention cost by $\mathcal{O}(T^2)$, where T is event volume time resolution, making it computationally prohibitive to train networks on bigger datasets. Due to these issues, we have yet to develop a generalizable event-based matching network with good zero-shot performance on unseen data. In this work, we seek to fill this missing link in event-based matching from two different directions: **architecture design** and **supervision**.

We design an event transformer that explicitly separates spatial and temporal aggregation to make multi-timescale tokens tractable at scale. Particularly, events present unique challenges that break the assumptions underlying 2D transformers for matching. First, events form an inherently 3D **spatiotemporal representation** rather than a fixed 2D grid. Second, motion profiles vary drastically across pixels, from slow drift to rapid edges, producing **nonuniform sampling** in time. To this end, we introduce a motion-aware and

separable feature encoder layer, which computes time and space feature aggregation efficiently. While naive multi-scale motion-aware layers would substantially increase the inference cost of transformers due to the large number of tokens, our proposed architecture mitigates this by decoupling spatial and temporal aggregation in a computationally efficient manner. Concretely, we use separable space and time attention to reduce the attention cost from $\mathcal{O}((THW)^2)$ to $\mathcal{O}(T(HW)^2 + HW\,T^2)$, which is the key enabling factor for training on millions of wide-baseline pairs rather than the narrow-baseline regimes used previously. In addition, we introduce an adaptive sparsity-aware event token selection module (SETS) that adaptively prunes redundant tokens for matching. By allocating computation only to informative spatiotemporal regions and time steps, SETS further reduces the effective cost of multi-timescale processing while preserving matching accuracy.

On the supervision side, we contribute two new datasets that cover a much wider range of view changes for event-based matching: a synthetic **E-MegaDepth** dataset and a real **Event Cross Matching (ECM)** dataset. E-MegaDepth contains simulated events generated from the large MegaDepth [42] dataset, featuring extensive, wide-baseline view changes and motion profiles. For real-world evaluation, we carefully designed a hardware-synchronized hetero-stereo setup that captures images and events simultaneously. Our ECM dataset enables real-world evaluation by providing synchronized images and event streams with dense and precise correspondence annotations across wide baselines. Using SOTA multi-view geometric foundation models, we generate point maps of higher density and precision compared to classical structure-from-motion solutions.

Finally, we trained our model using the combined large-scale dataset, achieving state-of-the-art results in semi-dense matching and camera pose estimation for both in-domain data and unseen test in-the-wild data without any fine-tuning. Our main contributions are as follows:

- **Zero-shot wide-baseline event matching**. We formalize the setting where viewpoint and motion profiles vary jointly, and propose a single matcher that generalizes to unseen datasets without test-time tuning, supporting both event-to-event and event-to-image correspondence.
- **Efficient multi-timescale event aggregation**. We introduce a motion-robust attention backbone that learns multi-timescale features via separable spatial–temporal aggregation, and further propose sparsity-aware event token selection to adaptively prune redundant tokens, resulting in efficient spatiotemporal transformers for events.
- **Wide-baseline supervision for events**. We release two complementary datasets, **E-MegaDepth** (large-scale synthetic wide-baseline pairs with diverse motions) and **ECM** (real hetero-stereo event–image data with dense correspondences), enabling scalable training and rigorous evaluation.

2 Related Work

Image-based Matching. Traditional image matching follows a detect-describe-match pipeline, where salient keypoints are extracted and encoded by handcrafted descriptors [6, 47, 62]. Their invariance abilities made them the backbone of early SfM [64] and SLAM systems [12], but they suffer from performance degradation in low-texture and fast motion scenes. Learning-based methods improved this pipeline through data-driven descriptors and joint detection-description, as shown in LIFT [78]. SuperPoint [16] popularized self-supervised keypoint learning, followed by [18, 60, 67], which enhanced feature confidence and distinctiveness. Later correspondence estimation models like SuperGlue [63], LightGlue [43], and Gluestick [54] leverage attention or graph reasoning to enforce geometric consistency, yet these detector-dependent approaches remain sparse and struggle in textureless or repetitive regions.

Detector-free methods establish correspondences directly from dense feature maps without keypoint detection. LoFTR [66] pioneered a coarse-to-fine transformer for semi-dense matching, showing strong robustness in low-texture areas. Later variants such as [41, 69, 71] improved efficiency and multi-scale consistency through local correlation and lightweight architectures. Beyond semi-dense matching, dense methods such as DKM [19], COTR [37], and RoMA [20] achieve pixel-level correspondence via global correlation and transformer reasoning across large viewpoints. These methods bridge image matching with optical flow and dense reconstruction, offering fine-grained details but having high computational cost. More recently, foundation models such as VGGT [68] have been developed to provide a unified approach to 3D prediction tasks. Our approach follows a semi-dense design that balances dense matches with high efficiency and competitive performance.

Event-based Matching. Event-based methods follow a similar trajectory as their image counterparts. Early studies adapted classical descriptors such as corners [3, 36, 48, 81], keypoints [2, 33, 40], and edges [10, 35] to events. However, these handcrafted features remain highly noise-sensitive and lack repeatable correspondences over wide baselines. To better exploit spatio-temporal information, subsequent works encode events into grid-like or volumetric structures that preserve temporal details, including time surfaces [7, 85], event volumes [24, 30, 56, 73, 84–86], and feature fields [15, 61]. Such representations laid the groundwork for downstream learning-based detection and matching models. Building on these, recent methods have directly learned local features for event-based correspondence. EventPoint [34] learns both keypoint detection and dense descriptors through spatio-temporal consistency, achieving self-supervised training without human annotations. SD2Event [22] further improved this framework by introducing a Contextual Feature Descriptor module to model long-range dependencies and Dynamic Keypoint Detector Learning module to adaptively detect keypoint. Zhu *et al.* [87] integrates multi-level temporal pyramids to enhance descriptor consistency and scale robustness in high-speed scenarios. In a

more recent work, SuperEvent [11] leverages image-based detectors on events to achieve robust detection and description under complex noise, and demonstrates strong performance in an event-based SLAM system.

Cross-Modal Matching. With the growing adoption of hybrid stereo setups that combine event and RGB cameras, event-to-image matching has become crucial for multimodal fusion and 3D perception [72]. Hybrid systems benefit from the complementary properties of both sensors–dense spatial information from frames and high temporal resolution from events. EKLT [25] explored event-RGB alignment by applying the classical KLT tracker on both events and frames. Beyond hybrid setups, event-based stereo matching itself has seen significant progress with improved disparity estimation [44, 55, 82] and spatial alignment in visual odometry [52]. More recent works [14, 38, 70] focused on direct event-image correspondence by jointly optimizing cross-modal features between events and frames. Later works [46, 80] extend these directions to zero-shot or wide-baseline configurations. At a broader scale, cross-modal matching frameworks, such as MINIMA [59] and MatchAnything [31], aim to learn foundational correspondence for diverse vision modalities.

3 Preliminaries

Event Camera Data. Event cameras asynchronously capture per-pixel changes in logarithmic brightness. When the change in a pixel exceeds a contrast threshold C, the sensor outputs an event $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) denotes the pixel coordinate, t_i is the timestamp and $p_i \in \{+1, -1\}$ indicates the polarity of the brightness change. The resulting event stream $E = \{e_i\}_{i=1}^{N}$ encodes fine-grained motion and texture of the scene that are useful for robot navigation.

Problem Formulation. We formally define the semi-dense event matching problem. Let $E^{\mathcal{A}}$ and $E^{\mathcal{B}}$ denote two event streams from viewpoints $\mathcal{A}$ and $\mathcal{B}$, respectively, which may differ by arbitrary baselines and motion profiles. Our task is to estimate a semi-dense correspondence field over this domain:

$$\mathcal{M}, \mathcal{P} = f_\theta(E^{\mathcal{A}}, E^{\mathcal{B}}), \tag{1}$$

where f_θ is a learnable matching function parameterized by network weights θ, $\mathcal{P} \in [0,1]^{H \times W}$ is the matching score map, and $\mathcal{M}(x, y) = (x', y')$ is the dense correspondence map. Each valid location in view $\mathcal{A}$ where $\mathcal{P}(x, y) > \omega$ is mapped to its corresponding coordinate (x', y') in view $\mathcal{B}$.

4 Method

Our method integrates event encoding, temporal aggregation, and dense matching into a unified, motion-robust end-to-end framework capable of predicting dense feature matches for both event-to-event and event-to-image matching tasks.

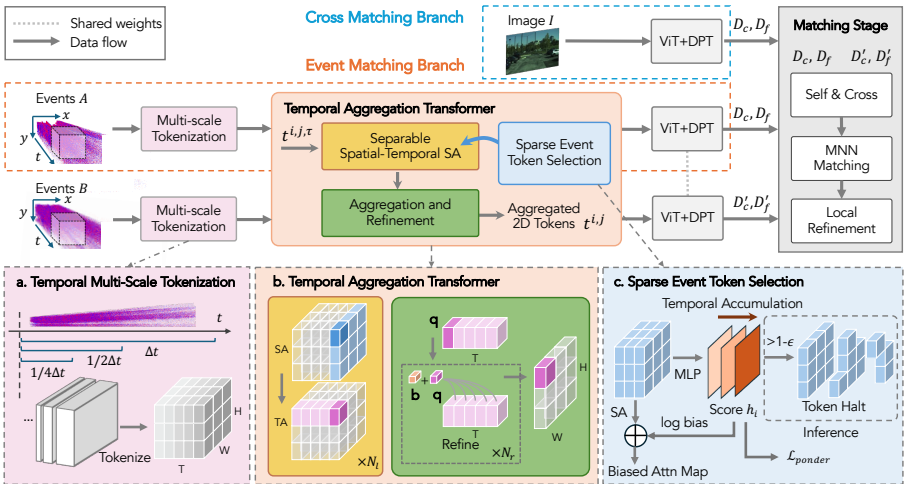

Fig. 2: Pipeline. Event slices are binned to multi-scale voxel preserving rich temporal information (Sec. 4.1). The voxel inputs are tokenized and processed by a separable Temporal Aggregation Transformer module (Sec. 4.2), and then with a sparsity-aware token selection (Sec. 4.3) module. Multi-scale and dense features from ViT followed by DPT are used for iterative matching in the final MNN matching module (Sec. 4.4).

4.1 Multi-timescale Event Representation

To use scalable architectures for event data, such as vision transformers, the first step is to design compatible input representations that preserve the rich temporal and spatial information. Early works convert event streams into 2D image-like tensors, such as event frames and time surfaces [7, 24]. However, rich temporal information is lost when events are squashed into a single surface [85]. Subsequently, event voxel representations [5, 85] emerged as a balanced solution between temporal resolution and computational efficiency.

We construct an event voxel logarithmically windowed in time [11, 29], as illustrated in Fig. 2. Specifically, an event stream is divided into B temporal bins, each representing a different accumulation time. Given N events $\{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, the voxel volume is defined as

$$V(c, x, y) = \sum_i p_i \, k_b(x - x_i) \, k_b(y - y_i) \, \mathbb{I}(c > c_i^*) \,, \tag{2}$$

where $c_i^* = \max(0, B - 1 + \log_2(\frac{t_i - t_1}{t_N - t_1}))$ and $k_b(a) = \max(0, 1 - |a|)$ denote a bilinear interpolation kernel, $\mathbb{I}(\cdot)$ is the indicator function, and c_i^* is the scaled event timestamp. After voxelization, the representation is tokenized into a set of tokens $\mathbf{t} \in \mathbb{R}^{T \times H \times W \times D}$ that retain temporal information. However, processing these multi-timescale temporal tokens effectively for dense correspondence requires a specialized architecture. To this end, we introduce a novel transformer that adaptively selects features across scales for high-quality matching.

Table 1: Quantitative comparison for event-to-event and event-to-image matching on ECM and M3ED. Red indicates the top performance, while orange and yellow denote the second and third, respectively.

	ECM Dataset				M3ED Dataset [13]			
	AUC@5°	AUC@10°	AUC@20°	Prec(%)	AUC@5°	AUC@10°	AUC@20°	Prec(%)
Event-to-Event Matching								
M-A (Events)	20.69	36.27	51.54	45.26	24.89	38.35	50.64	47.59
M-A (E2V)	41.68	59.71	73.01	50.05	38.04	50.69	60.47	46.79
VGGT (Events)	1.86	6.74	17.45	12.06	15.90	24.24	33.70	20.20
VGGT (E2V)	31.44	52.03	68.40	43.41	28.99	42.25	52.46	36.29
SuperEvent	11.40	22.12	34.24	33.93	18.07	29.34	40.65	39.43
Ours	**54.61**	**72.24**	**82.67**	**68.90**	**52.99**	**66.69**	**76.40**	**67.40**
Event-to-Image Matching								
M-A (Events)	23.60	38.93	52.99	41.53	26.35	38.92	50.22	43.01
M-A (E2V)	41.70	59.60	72.09	49.68	31.92	45.88	56.92	39.07
VGGT (Events)	2.18	6.32	13.17	8.29	4.26	8.47	14.23	9.34
VGGT (E2V)	18.10	34.25	50.75	30.94	13.17	21.70	29.85	21.02
Ours	**48.58**	**66.27**	**78.30**	**60.73**	**54.97**	**68.03**	**76.92**	**62.33**

Table 2: Pose estimation on EDS [32]. AUC is calculated with rotation errors.

Models	AUC@5°	AUC@10°	AUC@20°
RATE† [36]	2.1	5.1	10.3
EventPoint† [34]	1.6	2.8	5.2
Superevent [11]	25.4	37.5	49.0
Ours	**40.4**	**56.2**	**68.8**

4.2 Temporal Aggregation Transformer

The duration of an event sequence is often treated as a hyperparameter in event-based neural network design, affecting both training and inference performance. Short intervals preserve sharp spatial details but result in sparse and noisy representations, whereas longer intervals produce denser outputs at the cost of motion blur. While prior work [51] has proposed an auxiliary convolutional network, Concentrate Net, to recover a visually consistent 2D tensor, the temporal information is discarded due to averaging over time. To resolve this fundamental tradeoff, we introduce **Temporal Aggregation Transformer (TAg)**, which performs temporal attention and feature fusion within an end-to-end framework and directly learns from the training objective. Our proposed module has two stages: Separable Spatial-temporal Attention and Temporal Aggregation Refinement.

Separable Spatial-Temporal Attention. Inspired by recent developments in video vision transformers [4, 8, 28], we first introduce a separable attention module that alternates between spatial and temporal dimensions of the input tokens $\mathbf{t} \in \mathbb{R}^{T\times H\times W\times D}$. Specifically, the spatial self-attention operates on tokens $\mathbf{t}^{\tau} \in \mathbb{R}^{H\times W\times D}$ within each temporal bin τ, while the temporal self-attention

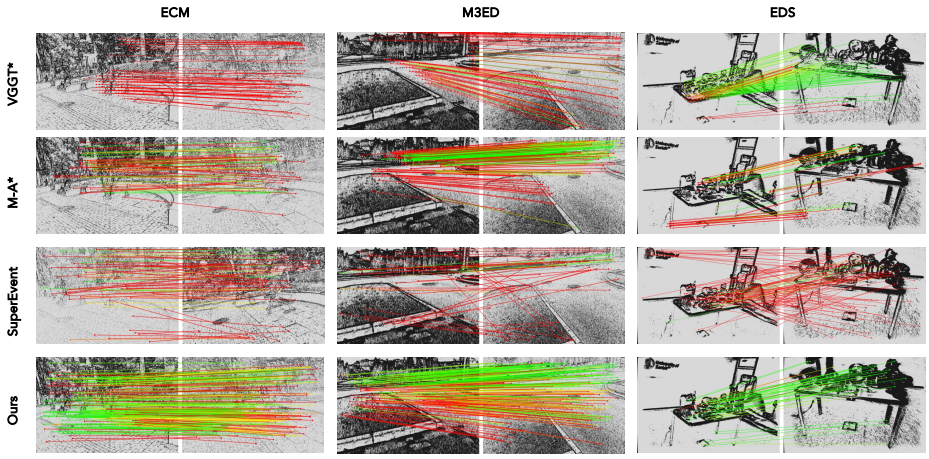

Fig. 3: Qualitative results of Match-Any-Events. We compared with VGGT [68], MatchAnything [31] and SuperEvent [11]. Match-Any-Events computes denser matches and achieves lower pose estimation errors. Green and Red indicate inliers and outliers for pose estimation respectively. Methods marked with $\star$ use event frames as input.

attends to tokens $\mathbf{t}_n \in \mathbb{R}^{T\times D}$ across the temporal dimension, where n is the spatial index. Intuitively, this encourages each embedding to gather information from its spatio-temporal neighbors, making the subsequent aggregation more robust. This design enables channel-wise and temporal feature fusion while reducing the computational cost from $O((THW)^2)$ to $O(T(HW)^2 + HWT^2)$. We interleave spatial and temporal attention for N_l times for fine feature extraction.

Temporal Aggregation and Refinement. To address event-specific problems like motion variance features under different binning window sizes, we introduce an event-based component to integrate multi-temporal-scale event maps into a canonical motion-invariant representation. Specifically, the second stage consolidates the information into a single feature map by querying the tokens in the first bin $\mathbf{t}^{\tau=0} \in \mathbb{R}^{H\times W\times D}$, which has the finest temporal resolution. The keys and values come from the remaining temporal bins within the same spatial indices. We directly query features from the bin with the finest temporal resolution, which contains sharp spatial details.

A learnable bias $\mathbf{b} \in \mathbb{R}^D$ is added to the query to enhance feature quality when the finest bin contains too little information. Due to the attention-based aggregation, the model remains robust to the input event interval, as shown in our ablation experiments. To better understand TAg, we visualize the **attention weights** of the first aggregation attention layer in Fig. 4. The model adaptively attends to fine-scale temporal bins in regions with fast motions, while shifting focus to coarse-scale temporal bins in regions with slow motions.

4.3 Sparsity-aware Event Token Selection (SETS)

Although the separable attention design in Fig. 4 significantly reduces the computational cost, our further analysis shows that the dense spatial attention introduces most of the FLOPs. Specifically, computing spatial attention at every temporal resolution yields a computational complexity of $O(T(HW)^2)$. Inspired by ACT [27], we propose **Sparsity-aware Event Token Selection (SETS)**, an adaptive token selection module to prune uninformative event tokens due to the inherent sparse nature of events.

Assuming the average spatial sparsity of event data is $\alpha \in (0, 1]$, the computational complexity can be optimized to $O(T\alpha^2(HW)^2)$. Unlike previous token pruning techniques that rely on fixed, sub-optimal reduction ratios [1, 57], SETS learns the sparsity structure of the data without needing task-specific prior knowledge. For each token $t_n^\tau \in \mathbb{R}^{T\times(HW)\times C}$ with spatial-temporal information, we predict a halting score h_n^τ at each temporal step τ through an MLP followed by a sigmoid. Scores are accumulated over temporal steps until they reach a threshold of $1-\epsilon$ at which point token processing halts. The variable N_n tracks the number of steps taken prior to halting. To ensure the mechanism remains completely differentiable and that the halting probabilities sum exactly to 1, we replace the final halting score with a remainder term $R_n = 1 - \sum_{i=1}^{N_n-1} h_n^i$.

We then define the ponder loss L_{ponder}, combining the remainder term R_n and a smoothing constant N_n, which encourages early halting and penalizes redundant temporal processing: $L_{ponder} = \frac{1}{HW}\sum_{n=1}^{HW}(N_n + R_n)$. Rather than formulating the neural module as a mean-field model and averaging the outputs across layers in prior work [27, 79], we directly inject the accumulated halting score as a bias term into the spatial attention weights in each temporal resolution:

$$\text{bias}_n^\tau = \begin{cases} \log\left(1 - \sum_{i=1}^{\tau-1} h_n^i\right), & \tau \le N_n \\ -\infty, & N_n < \tau \le T \end{cases} \tag{3}$$

We compute a bias map from the bias term bias_n^τ and inject it into the raw spatial attention map. By introducing this bias, tokens with high halting scores are actively suppressed during the attention operation. Consequently, the combination of the matching loss and ponder loss acts as a push-pull mechanism, effectively balancing predictive performance with computational efficiency.

4.4 Matching Stage

We address the semi-dense matching problem by progressively refining correspondences in a coarse-to-fine manner. Cross and self attentions are applied between two coarse feature maps (Stride=14) in an alternating fashion. The coarse features are correlated to build a score matrix S, on which dual-softmax is applied to obtain a matching probability P. The matching correspondence

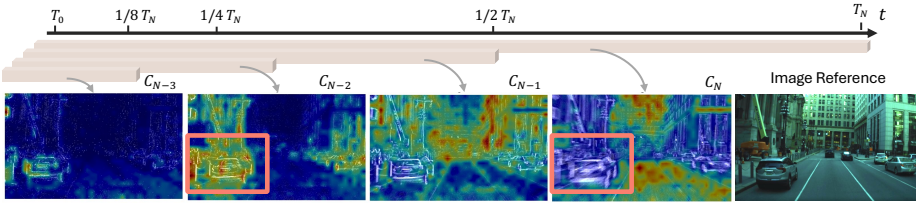

Fig. 4: Visualize Temporal Attention Weights. We visualize the attention weights of different time scales for the Aggregation and Refinement module. **Red** indicates high weight, while **blue** indicates low weight. The network attends to texture-rich regions at short temporal scales, and focuses on low-texture areas at longer temporal scales.

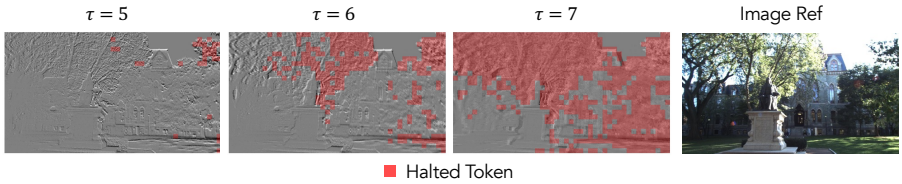

Fig. 5: Visualizing Sparsity-Aware Token Selection. We highlight the halted tokens at each temporal step using red patches. As the temporal information accumulates, an increasing number of tokens in spatially blurry regions are halted.

indices (x_i, y_j) are selected using Mutual Nearest Neighbors (MNN):

$$(x_i, y_j) = (\arg\max_x S(x, y_j), \arg\max_y S(x_i, y)) \tag{4}$$

The indices from coarse matching are used to crop fine features into local patches, where fine correspondences are determined with MNN similar to coarse features. Finally, a 3×3 window is cropped, and the predictions are refined by computing the expected values over the local region.

Loss functions. The coarse-level matching loss is formulated by minimizing the cross-entropy loss over the ground-truth matching matrix M_{gt}:

$$L_c = -\frac{1}{|M_c^{gt}|} \sum M_c^{gt} \log P_c. \tag{5}$$

Similarly, the fine-level loss is also supervised on the correlated matching matrix in fine scale: $L_f = -\frac{1}{|M_f^{gt}|} \sum M_f^{gt} \log P_f$. We use a refinement stage to generate sub-pixel level prediction. The refinement prediction is supervised by L_l, which is defined as an $\mathcal{L}_2$ loss between the predicted coordinate and ground-truth coordinate. The final loss consists of the losses for the coarse-level and the fine-level: $L = L_c + \alpha L_f + \beta L_l + \gamma L_{ponder}$.

5 E-MegaDepth and ECM Datasets

E-MegaDepth builds on MegaDepth [42], a large-scale internet photo collections dataset with poses and dense depth annotations. In contrast to the predom-

inantly driving oriented content of existing event datasets, MegaDepth offers rich diversity in viewpoint changes, making it suitable for evaluating wide-baseline matching. We construct a new synthetic event-based dataset, E-MegaDepth, by warping the images using a randomly sampled 6-DoF transformation. By interpolating these transformations, we synthesize high frame-rate video, which is further converted into event streams using Vid2E [23]. During event generation, we randomly select contrast sensitivity $C \sim \mathcal{U}(0.16, 0.34)$ following [39]. We generate approximately 3 million pairs of event streams for training.

Event Cross Matching Dataset (ECM). Unlike previous real event outdoor datasets collected under driving motions [49, 53, 83], which lack motion diversity and viewpoint variations that are important for generalizable matching. To address this gap, we collected a new dataset using a calibrated and synchronized multi-camera system including a Prophesee Gen4 camera (1280×720) and a Flir RGB camera (1920×1080). For annotation, we provide COLMAP [64] poses and bundle-adjusted VGGT [68] depth. Each scene is recorded with two different trajectories with diverse motion patterns and large viewpoint changes.

6 Experiments

In this section, we describe the experimental details, including datasets, baselines and metrics Sec. 6.1, and report evaluation results thoroughly. We focus on two tasks: event-to-event matching and event-to-image matching.

6.1 Data, Baselines, Metrics and Implementation

Datasets. We evaluate our method on M3ED [13], EDS [32], and ECM quantitatively. Please refer to Sec. 5 for details of ECM. The M3ED dataset provides multi-sensor recordings from ground, aerial, and legged robots equipped with event cameras, RGB cameras, LiDAR, and IMU. It provides synchronized event streams. To obtain ground-truth correspondences, we project 3D point clouds between frames using known sensor extrinsics and poses. We sample 20,000 pairs for training and 1,000 pairs for testing. EDS records indoor scenes under varying illumination using a beam splitter that captures synchronized events and images. Note that our model is not trained on either ECM or EDS, which enables real zero-shot evaluation.

Protocol. On **EDS**, we follow the evaluation setup of [11] and report the Area Under the Curve (AUC) of pose estimation errors at $(5°, 10°, 20°)$, using rotation error. On **ECM** and **M3ED**, we follow the evaluation setup of [66], where the pose error is defined as the maximum angular error in rotation and translation. To reduce randomness, we repeat the RANSAC five times and calculate the AUC over all errors. We also report matching precision following [63], defined as the average ratio of the correct matches to total matches, where a correct match has an epipolar distance below $1 \cdot 10^{-4}$ for outdoor scenes and $5 \cdot 10^{-4}$

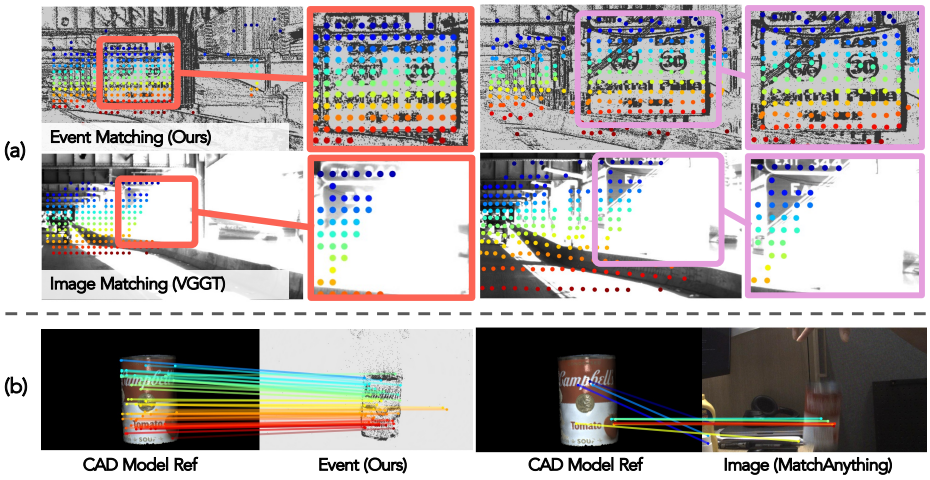

Fig. 6: Downstream tasks and applications. **(a)** Event and image matching in HDR scene. Event cameras provide distinguishable visual details, enabling high-confidence matching where standard images fail due to extreme lighting. (The example shown is from *schuylkill_tunnel* in M3ED [13]) **(b)** Downstream applications for cross modality matching: Pose estimation for fast-moving objects.

for indoor scenes. All images and events are resized so that the longer side is around 640 pixels. Unless otherwise specified, all models are evaluated using 128 ms inputs. Since the largest scale of SuperEvent input is 0.1 s, its input is also set to 128 ms. For MatchAnything and VGGT, we observed better performance with 30 ms input, which is therefore used as their evaluation interval.

Baseline Methods. SuperEvent [11] adopts a detector-based architecture that jointly predicts detections and descriptions, trained under supervision from image-based methods. MatchAnything [31] is trained on a large-scale self-labeled cross-modal dataset to generalize on unseen real-world cross-modality matching tasks. VGGT is a 1.2-billion-parameter model trained on a diverse dataset with 3D annotations, designed for geometric reasoning and uniform estimation of 3D properties within a scene. For methods originally trained on images, we report both zero–shot performance and the results using images reconstructed from events in order to provide a fair comparison.

Implementation Details. Our model is trained on a single NVIDIA H100 GPU (96G) for three days with a batch size of 16 and an input resolution of 560×336. We use the AdamW optimizer with a learning rate of $6 \cdot 10^{-4}$ and weight decay of 10^{-1}. The ViT backbone is initialized with DINO [65] pretrained weights, as we empirically observed that this accelerates convergence. All remaining modules are trained from scratch. We set $N_l = 2$ and $N_r = 4$. The full model runs at 49ms for a 350×630 event pair on an NVIDIA RTX 4080 GPU.

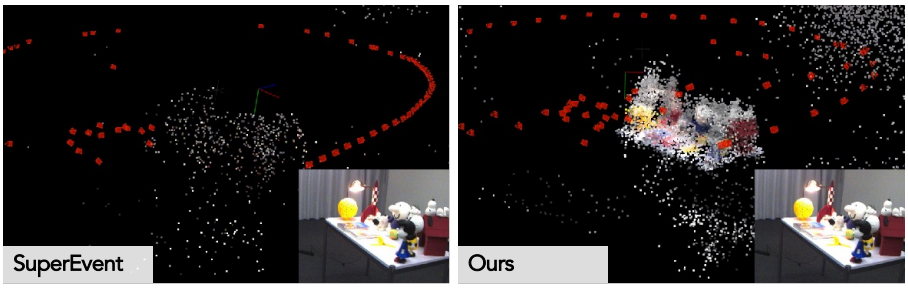

Fig. 7: Structure-from-Motion based on event matching. Left: SuperEvent [11]. **Right:** Ours. Our method produces more consistent camera poses and denser reconstruction. **Bottom Right of Each Image**: RGB frame for visualization reference.

Table 3: Ablation study on ECM.

Index	Ablation Models		AUC@5°	Prec(%)
I	Module	w/o TAg	50.12	67.18
II		I w/ Concentrate Net [51]	48.37	64.79
III		I w/o Multi-scale Input	39.86	66.87
IV	Dataset	Real Only(M3ED)	11.33	20.25
V		Syn Only(E-Mega)	52.63	68.04
VI	-	**Full**	**54.61**	**68.90**

6.2 Results

Results on ECM and M3ED. We evaluate the performance on our ECM and M3ED [13] test split, as is shown in Tab. 1. For ECM dataset, we build a sparse 3D reconstruction with incremental SfM pipeline in COLMAP [64] and calculate overlap scores based on the number of common 3D points between views. We generate 462 pairs with minimum overlap score of 0.05 for evaluation. To fairly evaluate MatchAnything(M-A) [31] and VGGT [68] which are not trained on events, we provide event frames (Events) and intensity reconstruction from E2VID [58]. Since our model does not rely on long-range temporal dependencies, we disable the recurrent connections in E2VID for consistency. Results in Tab. 1 show that our model learns robust matching predictions, offering **2.93 times** the AUC@5° than the state-of-the-art event-based baseline SuperEvent. We benchmarked computational efficiency on ECM. Our token selection mechanism reduces the FLOPs required for spatial attention operations by **21.5%** (48.87G vs 62.22G) with minimal impact on overall performance.

Results on EDS [32]. We evaluate our model on EDS using the test indices provided by [11]. SuperEvent achieves higher performance after correcting the intrinsic parameters compared to the results reported in the original paper. Therefore, we report these improved values in Tab. 2. Due to the missing source code, we are unable to report results for EventPoint [34]. Our method achieves

a **59%** improvement over the detector-based SuperEvent.

Downstream applications. We show the effectiveness of our matching on the *all_characters* sequence in EDS dataset using **only events**. The event stream is temporally downsampled into 66 segments, each selected to ensure sufficient translation. We perform exhaustive pairwise matching across these event segments and reconstruct both camera poses and 3D point clouds using COLMAP, as shown in Fig. 7. Compared to SuperEvent, Match-Any-Events produces significantly denser and clearer reconstructions and more consistent camera poses. Our matcher is capable of establishing reliable matches over large viewpoint changes, whereas SuperEvent struggles under large viewpoint changes and yields sparse keypoints. We also perform preliminary experiments on two downstream applications in Fig. 6 (a) and (b). Event cameras allow us to match more robustly in challenging scenarios than frame-based sensors due to dynamic range and motion blur.

Motion Robustness. We perform a thorough evaluation of motion robustness by measuring model performance while varying the input event window interval to simulate different motion speeds. Decreasing the interval below 20 ms results in a performance drop for all models, likely due to insufficient texture information. Quantitative results in Fig. 9 demonstrate that our model maintains consistent performance across a wide range of interval values. Additionally, incorporating Concentrate Net [51] enhances robustness but it slightly reduces peak performance. In contrast, Match-Any-Events exhibits almost no performance degradation as the interval increases.

Ablation Study. We evaluate four model variants in Tab. 3. Setup I is trained without Temporal Aggregation Transformer (TAg). Instead, the temporal dimension is flattened into channels by multiplying the input dimension of the embedding layer by N. Setup II does not have multi-temporal scale input. Setup III is built upon setup II by replacing TAg with Concentrate Net [51], a CNN that aggregates 3D volume into a 2D representation. We evaluated event-matching models trained on synthetic and real datasets in setup IV and V. Models trained on M3ED show limited generalization, likely because driving scenes provide relatively constrained motion diversity. Setup VI is our full model.

Event-to-Image Matching. We further evaluate cross-modal matching on ECM and M3ED, both of which provide RGB images and events. As indicated in Tab. 1, MatchAnything [31] achieves decent performance in this task, while VGGT struggles to adapt to reconstructed images from events. and performs poorly on both datasets. In contrast, our method consistently outperforms all baselines, surpassing MatchAnything by about 7 on AUC@5° on ECM and over 20 on AUC@5° on M3ED. Qualitative comparisons in Fig. 8 further demonstrate that our method produces more accurate correspondences across modalities.

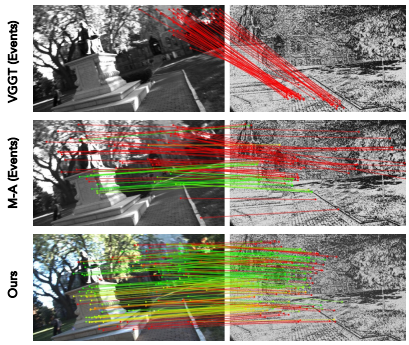

Fig. 8: Events-to-Image Matching. Match-Any-Events can match across wide baselines between images and events.

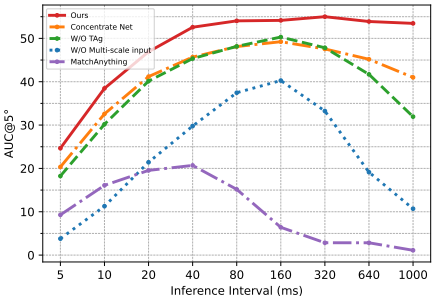

Fig. 9: Motion robustness study. Four variants of Match-Any-Events and MatchAnything [31] are evaluated on ECM under a controlled interval setting, and the AUC@5° results are reported.

6.3 Analysis

Our experiments show that the margin between our model and SuperEvent [11] is larger on ECM than on M3ED [13]. M3ED offers pairs from adjacent views, whereas our benchmark samples views randomly, resulting in larger viewpoint changes and discontinuous motion. The gain comes from our multi-view synthetic training set, which provides diverse motions and long-range correspondence supervision. The Temporal Aggregation module further prevents overfitting to specific motion profiles and strengthens learning from synthetic data. We also find that VGGT (Events) [68] performs better on the indoor EDS dataset [32], likely because indoor scenes contain structured geometry, such as lines and planes, that benefits the three dimensional reasoning typical of foundation models.

7 Discussion

We present Match-Any-Events, a state-of-the-art generalizable matching model for event data. Our approach addresses two fundamental issues in event-based matching: 1) we introduce an efficient and scalable spatiotemporal transformer and a sparsity-aware event token selection module, and 2) we curate a large-scale synthetic dataset E-MegaDepth for wide-baseline matching and collect Event Cross Matching, a real-world challenging hetero-stereo dataset. By combining an efficient architecture and diverse data sources, we develop a generalizable matching model that handles wide baselines, motion variations, and cross-modality correspondences. Match-Any-Events achieves state-of-the-art performance against all previous event matching methods, beating prior art by 37.7% in zero-shot performance on unseen datasets.

Acknowledgment. The financial support through the grants NSF FRR 2220868, NSF IIS 2212433, and ONR N00014-22-1-2677 is gratefully acknowledged. This work was also supported in part by funding from the Johns Hopkins Data Science and AI Institute.

References

1. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9392–9401 (2025)
2. Alzugaray, I., Chli, M.: HASTE: multi-hypothesis asynchronous speeded-up tracking of events. In: 31st British Machine Vision Conference 2020, BMVC 2020. BMVA Press (2020), https://www.bmvc2020-conference.com/assets/papers/0744.pdf
3. Alzugaray, I., Chli, M.: Asynchronous corner detection and tracking for event cameras in real time. IEEE Robotics and Automation Letters **3**(4), 3177–3184 (2018)
4. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
5. Baldwin, R.W., Liu, R., Almatrafi, M., Asari, V., Hirakawa, K.: Time-ordered recent event (tore) volumes for event cameras. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2519–2532 (2022)
6. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Eur. Conf. Comput. Vis. pp. 404–417. Springer (2006)
7. Benosman, R., Clercq, C., Lagorce, X., Ieng, S.H., Bartolozzi, C.: Event-based visual flow. IEEE transactions on neural networks and learning systems **25**(2), 407–417 (2013)
8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: International Conference on Machine Learning (ICML). pp. 813–824 (2021)
9. Born, R.T., Bradley, D.C.: Structure and function of visual area mt. Annu. Rev. Neurosci. **28**(1), 157–189 (2005)
10. Brändli, C., Strubel, J., Keller, S., Scaramuzza, D., Delbruck, T.: Elised—an event-based line segment detector. In: 2016 Second International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP). pp. 1–7. IEEE (2016)
11. Burkhardt, Y., Schaefer, S., Leutenegger, S.: Superevent: Cross-modal learning of event-based keypoint detection. arXiv preprint arXiv:2504.00139 (2025)
12. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE transactions on robotics **37**(6), 1874–1890 (2021)
13. Chaney, K., Cladera, F., Wang, Z., Bisulco, A., Hsieh, M.A., Korpela, C., Kumar, V., Taylor, C.J., Daniilidis, K.: M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4016–4023 (2023)
14. Cho, H., Yoon, K.J.: Selection and cross similarity for event-image deep stereo. In: Eur. Conf. Comput. Vis. pp. 470–486. Springer (2022)
15. Das, R., Daniilidis, K., Chaudhari, P.: Fast feature field (F^3): A predictive representation of events. arXiv preprint arXiv:2509.25146 (2025)
16. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 224–236 (2018)
17. DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron **73**(3), 415–434 (2012)
18. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8092–8101 (2019)

19. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: Dkm: Dense kernelized feature matching for geometry estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17765–17775 (2023)
20. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19790–19800 (2024)
21. Epstein, R., Kanwisher, N.: A cortical representation of the local visual environment. Nature **392**(6676), 598–601 (1998)
22. Gao, Y., Zhu, Y., Li, X., Du, Y., Zhang, T.: Sd2event: self-supervised learning of dynamic detectors and contextual descriptors for event cameras. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3055–3064 (2024)
23. Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Video to events: Recycling video datasets for event cameras. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3586–3595 (2020)
24. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Int. Conf. Comput. Vis. pp. 5633–5643 (2019)
25. Gehrig, D., Rebecq, H., Gallego, G., Scaramuzza, D.: Eklt: Asynchronous photometric feature tracking using events and frames. Int. J. Comput. Vis. **128**(3), 601–618 (2020)
26. Gibson, J.J.: The perception of the visual world. Houghton Mifflin (1950)
27. Graves, A.: Adaptive computation time for recurrent neural networks. arXiv preprint arXiv:1603.08983 (2016)
28. Gritsenko, A.A., Xiong, X., Djolonga, J., Dehghani, M., Sun, C., Lucic, M., Schmid, C., Arnab, A.: End-to-end spatio-temporal action localisation with video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18373–18383 (2024)
29. Hamann, F., Gehrig, D., Febryanto, F., Daniilidis, K., Gallego, G.: Etap: Event-based tracking of any point. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 27186–27196 (2025)
30. Hamann, F., Wang, Z., Asmanis, I., Chaney, K., Gallego, G., Daniilidis, K.: Motion-prior contrast maximization for dense continuous-time motion estimation. In: European Conference on Computer Vision. pp. 18–37. Springer (2024)
31. He, X., Yu, H., Peng, S., Tan, D., Shen, Z., Bao, H., Zhou, X.: Matchanything: Universal cross-modality image matching with large-scale pre-training. arXiv preprint arXiv:2501.07556 (2025)
32. Hidalgo-Carrió, J., Gallego, G., Scaramuzza, D.: Event-aided direct sparse odometry. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5781–5790 (2022)
33. Hu, S., Kim, Y., Lim, H., Lee, A.J., Myung, H.: ecdt: Event clustering for simultaneous feature detection and tracking. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3808–3815. IEEE (2022)
34. Huang, Z., Sun, L., Zhao, C., Li, S., Su, S.: Eventpoint: Self-supervised interest point detection and description for event-based camera. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5396–5405 (2023)
35. Ikura, M., Glover, A., Mizuno, M., Bartolozzi, C.: Lattice-allocated real-time line segment feature detection and tracking using only an event-based camera. In: Int. Conf. Comput. Vis. pp. 4645–4654 (2025)
36. Ikura, M., Le Gentil, C., Müller, M.G., Schuler, F., Yamashita, A., Stürzl, W.: Rate: Real-time asynchronous feature tracking with event cameras. In: 2024 IEEE/RSJ

International Conference on Intelligent Robots and Systems (IROS). pp. 11662–11669. IEEE (2024)
37. Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: Cotr: Correspondence transformer for matching across images. In: Int. Conf. Comput. Vis. pp. 6207–6217 (2021)
38. Kim, H., Lee, S., Kim, J., Kim, H.J.: Real-time hetero-stereo matching for event and frame camera with aligned events using maximum shift distance. IEEE Robotics and Automation Letters **8**(1), 416–423 (2022)
39. Klenk, S., Motzet, M., Koestler, L., Cremers, D.: Deep event visual odometry. In: 2024 International conference on 3D vision (3DV). pp. 739–749. IEEE (2024)
40. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: a hierarchy of event-based time-surfaces for pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(7), 1346–1359 (2016)
41. Li, X., Rao, T., Pan, C.: Edm: Efficient deep feature matching. arXiv preprint arXiv:2503.05122 (2025)
42. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2041–2050 (2018)
43. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Int. Conf. Comput. Vis. pp. 17627–17638 (2023)
44. Liu, P., Chen, G., Li, Z., Tang, H., Knoll, A.: Learning local event-based descriptor for patch-based stereo matching. In: International Conference on Robotics and Automation (ICRA). pp. 412–418. IEEE (2022)
45. Logothetis, N.K., Sheinberg, D.L.: Visual object recognition. Annual review of neuroscience **19**, 577–621 (1996)
46. Lou, H., Liang, J., Teng, M., Fan, B., Xu, Y., Shi, B.: Zero-shot event-intensity asymmetric stereo via visual prompting from image domain. Advances in Neural Information Processing Systems **37**, 13274–13301 (2024)
47. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
48. Mueggler, E., Bartolozzi, C., Scaramuzza, D.: Fast event-based corner detection. In: British Machine Vision Conference (BMVC) (2017)
49. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. The International journal of robotics research **36**(2), 142–149 (2017)
50. Murray, S.O., Wojciulik, E.: Attention increases neural selectivity in the human lateral occipital complex. Nature neuroscience **7**(1), 70–74 (2004)
51. Nam, Y., Mostafavi, M., Yoon, K.J., Choi, J.: Stereo depth from events cameras: Concentrate and focus on the future. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6114–6123 (2022)
52. Niu, J., Zhong, S., Lu, X., Shen, S., Gallego, G., Zhou, Y.: Esvo2: Direct visual-inertial odometry with stereo event cameras. IEEE Transactions on Robotics (2025)
53. Orchard, G., Jayawant, A., Cohen, G.K., Thakor, N.: Converting static image datasets to spiking neuromorphic datasets using saccades. Frontiers in neuroscience **9**, 437 (2015)
54. Pautrat, R., Suárez, I., Yu, Y., Pollefeys, M., Larsson, V.: Gluestick: Robust image matching by sticking points and lines together. In: Int. Conf. Comput. Vis. pp. 9706–9716 (2023)
55. Piatkowska, E., Kogler, J., Belbachir, N., Gelautz, M.: Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 53–60 (2017)

56. Qu, Q., Chen, X., Chung, Y.Y., Shen, Y.: Evrepsl: Event-stream representation via self-supervised learning for event-based vision. IEEE Trans. Image Process. (2024)
57. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)
58. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE Trans. Pattern Anal. Mach. Intell. **43**(6), 1964–1980 (2019)
59. Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X.: Minima: Modality invariant image matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23059–23068 (2025)
60. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2d2: Reliable and repeatable detector and descriptor. Adv. Neural Inform. Process. Syst. **32** (2019)
61. Rodriguez, L.G., Konrad, J., Drees, D., Risse, B.: S-rope: Spectral frame representation of periodic events. In: Eur. Conf. Comput. Vis. pp. 307–324. Springer (2024)
62. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Int. Conf. Comput. Vis. pp. 2564–2571. Ieee (2011)
63. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4938–4947 (2020)
64. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
65. 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)
66. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8922–8931 (2021)
67. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. Adv. Neural Inform. Process. Syst. **33**, 14254–14265 (2020)
68. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025)
69. Wang, Q., Zhang, J., Yang, K., Peng, K., Stiefelhagen, R.: Matchformer: Interleaving attention in transformers for feature matching. In: ACCV. pp. 2746–2762 (2022)
70. Wang, X., Yu, H., Yu, L., Yang, W., Xia, G.S.: Towards robust keypoint detection and tracking: A fusion approach with event-aligned image features. IEEE Robotics and Automation Letters (2024)
71. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21666–21675 (2024)
72. Wang, Z., Pan, L., Ng, Y., Zhuang, Z., Mahony, R.: Stereo hybrid event-frame (shef) cameras for 3d perception. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9758–9764. IEEE (2021)
73. Wang, Z., Chaney, K., Daniilidis, K.: EvAC3D: From event-based apparent contours to 3D models via continuous visual hulls. In: Eur. Conf. Comput. Vis. pp. 284–299 (2022)

74. Wang, Z., Cladera, F., Bisulco, A., Lee, D., Taylor, C.J., Daniilidis, K., Hsieh, M.A., Lee, D.D., Isler, V.: EV-Catcher: High-speed object catching using low-latency event-based neural networks. IEEE Robotics and Automation Letters **7**(4), 8737–8744 (2022)
75. Wang, Z., Hamann, F., Chaney, K., Jiang, W., Gallego, G., Daniilidis, K.: Event-based continuous color video decompression from single frames. arXiv preprint arXiv:2312.00113 (2023)
76. Wang, Z., Zhang, R., Liu, Z.Y., Wang, Y., Daniilidis, K.: Continuous-time human motion field from event cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11502–11512 (2025)
77. Wang, Z.C.: Beyond Frames: Learning to Perceive With Event-Based Vision. Ph.D. thesis, University of Pennsylvania (2025)
78. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: Eur. Conf. Comput. Vis. pp. 467–483. Springer (2016)
79. Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10809–10818 (2022)
80. Zhang, D., Ding, Q., Duan, P., Zhou, C., Shi, B.: Data association between event streams and intensity frames under diverse baselines. In: Eur. Conf. Comput. Vis. pp. 72–90. Springer (2022)
81. Zhu, A.Z., Atanasov, N., Daniilidis, K.: Event-based feature tracking with probabilistic data association. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 4465–4470. IEEE (2017)
82. Zhu, A.Z., Chen, Y., Daniilidis, K.: Realtime time synchronized event-based stereo. In: Eur. Conf. Comput. Vis. pp. 433–447 (2018)
83. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)
84. Zhu, A.Z., Wang, Z., Khant, K., Daniilidis, K.: EventGAN: Leveraging large scale image datasets for event cameras. In: IEEE International Conference on Computational Photography (ICCP). pp. 1–11 (2021)
85. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv preprint arXiv:1802.06898 (2018)
86. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 989–997 (2019)
87. Zhu, Y., Gao, Y., Ding, T., Liu, X., Yang, W., Zhang, T.: Spatio-temporal pyramid keypoint detection with event cameras. IEEE Transactions on Circuits and Systems for Video Technology (2025)