# ISTASTrack: Bridging ANN and SNN via ISTA Adapter for RGB-Event Tracking

Siying Liu, Zikai Wang, Hanle Zheng, Yifan Hu, Xilin Wang, Qingkai Yang, Jibin Wu, *Member, IEEE,* Hao Guo, Lei Deng, *Senior Member, IEEE,*

arXiv:2509.09977v1 [cs.CV] 12 Sep 2025

*Abstract*—RGB-Event tracking has become a promising trend in visual object tracking to leverage the complementary strengths of both RGB images and dynamic spike events for improved performance. However, existing artificial neural networks (ANNs) struggle to fully exploit the sparse and asynchronous nature of event streams. Recent efforts toward hybrid architectures combining ANNs and spiking neural networks (SNNs) have emerged as a promising solution in RGB-Event perception, yet effectively fusing features across heterogeneous paradigms remains a challenge. In this work, we propose ISTASTrack, the first transformer-based ANN-SNN hybrid Tracker equipped with ISTA adapters for RGB-Event tracking. The two-branch model employs a vision transformer to extract spatial context from RGB inputs and a spiking transformer to capture spatio-temporal dynamics from event streams. To bridge the modality and paradigm gap between ANN and SNN features, we systematically design a model-based ISTA adapter for bidirectional feature interaction between the two branches, derived from sparse representation theory by unfolding the iterative shrinkage thresholding algorithm. Additionally, we incorporate a temporal downsampling attention module within the adapter to align multi-step SNN features with single-step ANN features in the latent space, improving temporal fusion. Experimental results on RGB-Event tracking benchmarks, such as FE240hz, VisEvent, COESOT, and FELT, have demonstrated that ISTASTrack achieves state-of-the-art performance while maintaining high energy efficiency, highlighting the effectiveness and practicality of hybrid ANN-SNN designs for robust visual tracking. The code is publicly available at https://github.com/lsying009/ISTASTrack.git.

*Index Terms*—Hybrid neural networks, spiking neural networks, sparse representation, RGB-Event fusion, multimodal object tracking.
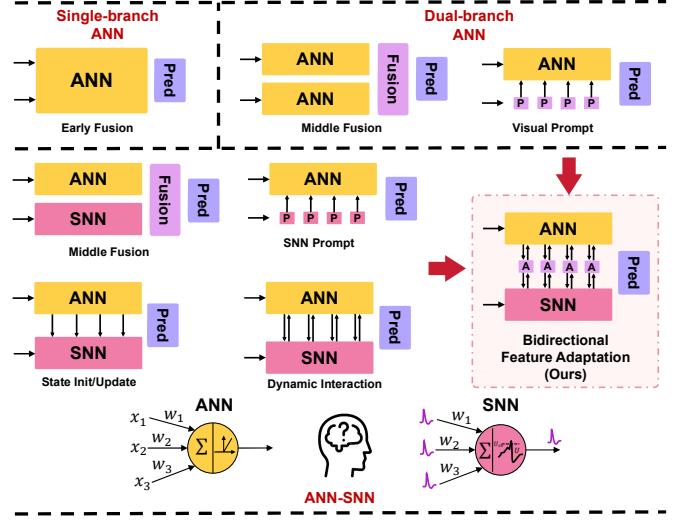
Fig. 1. ANN and hybrid ANN-SNN architecture with different fusion strategies in RGB-Event perception. Our ISTASTrack exploits bidirectional feature interaction between the ANN and SNN branches.

Siying Liu, Hanle Zheng, Yifan Hu, and Lei Deng are with the Center for Brain Inspired Computing Research (CBICR), Department of Precision Instrument, Tsinghua University, Beijing, China (e-mail: siying-liu@mail.tsinghua.edu.cn; zhl22@mails.tsinghua.edu.cn; huyf19@mails.tsinghua.edu.cn; leideng@mail.tsinghua.edu.cn).

Zikai Wang and Hao Guo are with the College of Computer Science and Technology, Taiyuan University of Technology, Shanxi, China (e-mail: zikaiwang@link.tyut.edu.cn; guohao@tyut.edu.cn).

Xilin Wang is with the Engineering Laboratory of Power Equipment Reliability in Complicated Coastal Environments, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China (e-mail: wang.xilin@sz.tsinghua.edu.cn).

Qingkai Yang is with the School of Automation, Beijing Institute of Technology, Beijing, China (e-mail: qingkai.yang@bit.edu.cn).

Jibin Wu is with the Department of Data Science and Artificial Intelligence and the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR (e-mail: jibin.wu@polyu.edu.hk).

## I. INTRODUCTION

Event cameras introduce a novel paradigm for visual perception by asynchronously detecting pixel-level intensity changes and generating dynamic spike events [1]. Their unique sensing mechanism offers advantages such as high temporal resolution, high dynamic range (HDR), and low power consumption, making them well suited for fast motion and extreme lighting conditions. However, event data lack the rich spatial details that traditional RGB cameras provide, such as texture and color. Therefore, RGB-Event fusion has shown great potential in advancing a wide range of computer vision tasks, such as object detection [2], visual tracking [3], and action recognition [4]. In this study, we focus specifically on visual object tracking (VOT), where effectively exploiting the complementary strengths of RGB and event modalities remains a key challenge.

Most existing RGB-Event tracking frameworks employ artificial neural networks (ANNs) to process both RGB and event modalities. As illustrated in Fig. 1, they can be broadly divided into single-branch and dual-branch architectures. Classic single-modality trackers such as DiMP [5] and OSTrack [6] can be extended to multi-modal settings through early fusion. CEUTrack [7] further adapts the transformer framework by embedding RGB-Event inputs into a shared token space. On the other hand, dual-branch models employ either symmetric

ANN backbones or modality-specific backbones, followed by dedicated middle fusion modules. Recent RGB-Event tracking approaches, including FENet [8], AFNet [9], CMT-MDNet [10], and CrossEI [11], incorporate cross-attention mechanisms to integrate multimodal features. With the success of vision transformer (ViT), RGB-X tracking has explored visual prompt tuning (VPT) by integrating auxiliary prompts into the RGB backbone [12]–[14]. However, most approaches focus on guiding RGB streams with auxiliary modalities, with only BAT [14] enabling bidirectional interaction to exploit complementary features.

Although ANNs are highly effective in extracting spatial correlations and global contextual information in RGB images, they are less effective for the sparse spatial structures and rich temporal dynamics of event data. To adapt to the frame-based architecture of ANNs, event streams are often converted into event frames aligned with the RGB frame rate, resulting in the loss of fine-grained temporal information. In contrast, spiking neural networks (SNNs) offer a more natural framework for temporal information processing through their intrinsic neural dynamics. This has led to an increasing interest in creating hybrid neural networks (HNNs) that integrate ANNs and SNNs, aiming to exploit both spatial features of ANNs and temporal dynamics of SNNs for more effective RGB-Event processing [15]–[19]. However, their application to object tracking remains relatively limited. MMHT [20] represents the first attempt at HNNs on VOT tasks by employing ResNet-based backbones and using transformer-based fusion modules for ANN-SNN feature fusion, but its performance is suboptimal. More recently, SNNPTrack [21] introduces event prompt using SNNs into a transformer-based RGB tracker. Despite these efforts, the tracking precision of current HNN-based approaches still lags behind mainstream ANN-based models.

Given the discrepancy between RGB-Event modalities and ANN-SNN paradigms, effective fusion strategies are crucial. As shown in Fig. 1, similar to ANN-based models, middle fusion [18], [20] and visual prompt [21] have been explored. Other works investigate initializing SNN states with ANN features [22] or employing interactive fusion during feature extraction [17], [19]. However, most methods still rely on attention-based fusion, which leads to high computational and memory cost and lacks of interpretability regarding how RGB and event information are jointly utilized. In recent years, sparse representation [23] has emerged as a model-based paradigm for interpretable multimodal alignment and fusion [24]–[26]. It facilitates establishing explicit cross-modal relationships by compactly decomposing multimodal data into shared and modality-specific components. Iterative sparse coding algorithms, such as the iterative shrinkage-thresholding algorithm (ISTA) [27], can be unfolded into lightweight network modules, bridging principled optimization with efficient deep architectures [28], [29]. However, this approach is usually built on convolutional neural network (CNN) architectures and focuses on pixel- or patch-level fusion, leaving their application to ANN-SNN hybrid networks and transformer-based frameworks largely unexplored. Inspired by these advances, we design a novel ISTA adapter that enables cross-modal interaction between ANN and SNN transformer encoders guided by sparse coding principles.

In this study, we propose ISTASTrack, an **A**NN-**S**NN hybrid **Track**er equipped with **ISTA** adapters for RGB-Event tracking. To the best of our knowledge, it is the first transformer-based ANN–SNN hybrid framework tailored to multimodal VOT, giving us valuable insights into hybrid network design and cross-paradigm feature fusion. The hybrid architecture incorporates a vision transformer branch to extract structural features from RGB inputs and a spiking transformer branch to exploit the spatio-temporal dynamics inherent in event data. To effectively leverage the complementary features of the ANN and SNN branches, we design a lightweight ISTA adapter that enables bidirectional feature adaptation and fusion. Specifically, the interaction between RGB and event features is formulated as a sparse coding problem, which can be solved using the ISTA algorithm. By applying the algorithm unfolding strategy, the iterative process of ISTA is transformed into trainable ISTA adapter modules. Furthermore, a temporal downsampling attention module (TDA) is integrated into the ISTA adapter to enhance temporal feature fusion by aligning multi-step SNN features with single-step ANN features in the latent space. Experimental results have demonstrated that our ISTASTrack consistently outperforms state-of-the-art trackers across multiple datasets, such as FE240hz, VisEvent, COESOT, and FELT, validating the effectiveness and efficiency of hybrid architecture and feature fusion strategies.

The main contributions of this work are summarized below:

1) We propose the first transformer-based ANN-SNN hybrid framework for RGB-Event tracking, which effectively leverages contextual and spatial information from RGB data and dynamic temporal cues from event data, achieving state-of-the-art performance with high efficiency.

2) A novel ISTA adapter is systematically developed based on the sparse representation theory, enabling bidirectional interactive feature adaptation and fusion between heterogeneous features while maintaining a lightweight and interpretable design. In addition, a TDA module is incorporated to facilitate semantic temporal alignment and fusion between SNN and ANN features.

3) ISTASTrack achieves state-of-the-art accuracy with low computational cost on extensive RGB-Event tracking benchmarks, including FE240hz, VisEvent, COESOT, and FELT. The comprehensive analyses further provide insightful guidance for the design of effective and efficient hybrid networks.

## II. RELATED WORKS

### A. ANN-based RGB-Event Object Tracking

VOT aims to locate a target in each video frame by matching a given template from the first frame with a search region in subsequent frames. ANN-based VOT methods have evolved from early Siamese neural networks (e.g., SiamFC [30] and SiamRPN [31]) and discriminative learning frameworks (e.g., ATOM [32], DiMP [5], PrDiMP [33], and KeepTrack [34]) to transformer-based architectures (e.g., STARK [35], SwinTrack [36], AiATrack [37], MixFormer [38], and OSTrack [6]), which have now become the mainstream due to their superior global modeling capacity and robust feature representation.

The development of RGB-Event VOT follows the progression of single-modal tracking, extending existing frameworks to incorporate multimodal fusion. Early studies typically employed CNN-based backbones like ResNet for independent feature extraction from each modality, followed by cross-modality fusion modules. For example, FENet [8] proposed a cross-domain attention scheme and an adaptive weighting strategy to dynamically balance modality contributions. AFNet [9] designed an alignment module to mitigate cross-style misalignment guided by event motion cues, alongside a fusion module to enhance salient features while suppressing noise. Wang *et al.* [10] incorporated a cross-modality transformer into the multi-domain network and ATOM to leverage complementary characteristics from the two modalities. Similarly, Han *et al.* [39] introduced self-attention modules to enhance the joint representation of RGB and event input. Recently, CrossEI [11] used explicit motion estimation and semantic modulation to align and integrate event-image data for dynamic scenes. However, the performance of these approaches remains limited, largely due to the representational constraints of the CNN-based backbones.

In recent years, one-stage transformer trackers like OS-Track [6] have rapidly advanced RGB–Event VOT. CEUTrack [7] proposed the first unified single-stage transformer-based framework by embedding multimodal inputs into a shared token space, which performs feature extraction, fusion, and interaction simultaneously. Building on this, AMTTrack [40] incorporated modern Hopfield layers to strengthen associative memory and effectively handle incomplete and uncertain multimodal information for robust long-term tracking. However, unified frameworks risk insufficient feature extraction when the discrepancy between modalities is large. To address relaxed registration for unaligned RGB and event pair, CRSOT [41] introduced probabilistic feature representation and modality uncertainty fusion into the OSTrack framework. TENet [42] proposed an event-specific backbone with multi-scale pooling to capture motion-aware features from sparse event data, followed by cross-attention interaction with the RGB branch.

In parallel with intricate fusion module design, another line of research focuses on lightweight adaptation. Approaches like ViPT [43], OneTracker [44], BAT [45], EMTrack [12] and UnTrack [13] use VPT technique by designing adapters to transform features from one modality into cross-modal prompts. These adapters are inserted between transformer blocks to enable latent space feature mapping. By exclusively tuning these lightweight adapters while keeping the pretrained backbone frozen, these methods achieve accurate dual-modal tracking with minimal training cost. Inspired by such design, we introduce interactive ISTA adapters; however, we do not freeze any parameters during training due to the large discrepancy between ANN and SNN features.

### B. Hybrid Neural Networks

Inspired by complementary learning mechanisms [46], [47] observed in biological brains, HNNs that integrate ANNs and SNNs have demonstrated advantages in perception, cognition, and learning by improving accuracy, robustness, and energy efficiency [48], [49]. For example, Zhao *et al.* [48] introduced a general HNN framework by proposing hybrid units that both integrate and decouple features from ANNs and SNNs, enhancing flexibility and computational efficiency. CH-HNN [50] mitigated catastrophic forgetting in continual learning by leveraging corticohippocampal-inspired circuits to enable energy-efficient, dual-memory representations. Additionally, hybrid spatiotemporal neural networks (HSTNNs) [51] improved adaptability across different performance metrics by flexibly adjusting the proportion of artificial and spiking neurons.

In multimodal computer vision tasks involving two distinct visual processing pathways, the core challenge of HNNs lies in designing fusion mechanisms that can effectively align and synchronize ANN and SNN features across both spatial and temporal domains. Recent works on RGB-Event tasks have explored HNNs in applications such as action recognition [17], video restoration [15], [16], object detection [18], [19] and tracking [20], [52]. Similar to the ANN-based paradigm, some approaches explore middle-stage fusion after separate feature extraction. For example, SC-Net [15] proposed a spiking convolutional architecture with a CNN-based spatial aggregation module for video restoration. Liu *et al.* [16] designed a motion-guided SNN-CNN hybrid encoder to fuse motion and background features for image deblurring. SSTFormer [18] and MMHT [20] employed transformer-based multimodal fusion modules to fuse ANN and SNN features. EOLO [53] introduced a symmetric fusion module to adaptively balance RGB and event cues for robust detection under varying lighting conditions.

However, treating ANN and SNN feature extraction separately limits the ability of middle-stage fusion to exploit their complementary strengths. To this end, some studies enhance the interactive mechanism between ANN and SNN branches during feature extraction. For example, STNet [52] leveraged an SNN to capture temporal cues while adaptively adjusting membrane thresholds guided by global spatial context from a transformer, enabling robust event tracking. Aydin *et al.* [22] used an auxiliary ANN running at a lower frequency to initialize SNN states, mitigating transient periods and state degradation. ReSpike [17] incorporated multi-scale cross-attention during ResNet extraction for action recognition. HDI-Former [19] proposed a biologically inspired hybrid dynamic interaction for effective cross-modal information exchange between transformer encoders, achieving high-accuracy object detection.

Research on HNNs for RGB-Event VOT is still in its early stages, with only a few dedicated studies, such as MMHT [20] and SNNPTrack [21]. MMHT employs simple spike-based AlexNet and ResNet as SNN and ANN backbones, followed by cross attention feature fusion, which restricts its performance. More recently, SNNPTrack [21] introduced SNN-based modules to extract temporal cues from events and progressively fuse them into transformer encoders. However, its uni-directional fusion design cannot fully exploit the complementary representations of both modalities. In contrast, our approach leverages transformers to capture rich contextual information and introduces a bidirectional feature adaptation

mechanism, enabling more effective and symmetric interaction between ANN and SNN branches.

### C. Sparse Coding Model for Multimodal Fusion

Sparse coding [23] is a classical representation learning framework that represents input signals as a sparse linear combination of atoms from an overcomplete dictionary. It offers a framework to capture the underlying structure or essential features of the data using a compact representation. In multimodal settings, it facilitates the separation of shared and modality-specific features, making it well suited for multimodal decomposition and fusion.

Traditional sparse coding problems are typically solved by iterative optimization algorithms such as the ISTA [27], approximate message passing (AMP) [54], and alternating direction method of multipliers (ADMM) [55]. With the rise of deep learning, algorithm unfolding (or unrolling) [28], [29] has emerged as an effective strategy to transform these iterative procedures to interpretable deep network architectures, embedding domain knowledge into learning. This model-driven approach has demonstrated success across various computational imaging tasks, including image restoration [56]–[58], super-resolution [59], [60], and image reconstruction [61].

In the context of multimodal fusion, many works design deep sparse coding models to extract task-specific features from each modality. For example, CU-Net [24] proposed a multimodal convolutional sparse coding model to split the common information shared among different modalities for image restoration and fusion. Similarly, Marivani *et al.* [25] used the method of multipliers to design CNNs. Deng *et al.*. [62] formulated multimodal image registration as a disentangled convolutional sparse coding model and unfolded its optimization into an interpretable network to separate registration-relevant and irrelevant features across modalities. Li *et al.* [26] proposed to learn sparse and discriminative multimodal features for finger recognition.

However, most existing works are built upon CNN architectures, focusing mainly on pixel- or patch-level reconstruction, while their integration into transformer-based multimodal fusion frameworks remains unexplored. To this end, we propose a novel ISTA adapter that introduces interaction mechanisms between ANN-SNN transformer encoders, enabling explicit cross-modal feature mapping in the transform domain guided by sparse coding principles.

### III. METHODOLOGY

In this section, we introduce the proposed ISTASTrack framework for RGB-Event VOT. We begin with an overview of the hybrid architecture in Sec. III-A. Next, Sec. III-B presents the transformer-based ANN and SNN backbones for feature extraction. In Sec. III-C, we design an ISTA adapter based on sparse representation and algorithm unfolding to promote heterogeneous feature adaptation across transformer encoders. Sec. III-D then describes a temporal downsampling attention module to align multi-step SNN features with single-step ANN features. Finally, Sec. III-E provides an analytical study of network computational cost.

### A. Overview of ISTATrack

Single VOT begins with a known target location in the first frame of a video and aims to localize the object in subsequent frames. Specifically, given the initial bounding box $B_0$ of the target object on the first RGB frame $I_0$, a tracking model $\mathcal{T}$ predicts the target position and size $B_i = (c_{x_i}, c_{y_i}, w_i, h_i)$ in the following frames $I_i$. At each step, the model receives a template image $\boldsymbol{Z}_{I_0} \in \mathbb{R}^{3 \times H_1 \times W_1}$ centered on the initial target and a search image $\boldsymbol{X}_{I_i} \in \mathbb{R}^{3 \times H_2 \times W_2}$ cropped around the previous prediction. The template is typically half the size of the search image in both height and width, where $H_1 = H_2/2$ and $W_1 = W_2/2$.

In the RGB-Event system, two asynchronous data streams are generated: a sequence of RGB frames $\boldsymbol{I}_i$ captured at time steps $t_i$, and a continuous stream of events $\{e_j, j = 1, 2, \cdots, N\}$. Events that occur between two consecutive frames $(t_i, t_{i+1})$ are accumulated and encoded into an event representation $\boldsymbol{E}_i$. In this work, we construct $\boldsymbol{E}_i$ as a stack of $T$ event frames for consistency with prior studies [9], [40], [63]. This results in a temporally aligned input pair $(\boldsymbol{I}_i, \boldsymbol{E}_i)$ for each time step. Based on this, the RGB tracking pipeline can be extended to support dual-modality inputs.

As shown in Fig. 2, the proposed RGB-Event VOT framework takes four inputs at each step: the RGB template $\boldsymbol{Z}_{I_0}$ and search image $\boldsymbol{X}_{I_i}$, together with the corresponding event template $\boldsymbol{Z}_{E_0} \in \mathbb{R}^{T \times 3 \times H_1 \times W_1}$ and search input $\boldsymbol{X}_{E_i} \in \mathbb{R}^{T \times 3 \times H_2 \times W_2}$. The predicted bounding box $B_i$ in frame $i$ is given by

$$\mathcal{B}_i = \mathcal{T}(\boldsymbol{Z}_{I_0}, \boldsymbol{Z}_{E_0}, \boldsymbol{X}_{I_i}, \boldsymbol{X}_{E_i}, B_0). \tag{1}$$

The architecture of ISTASTrack is depicted in Fig. 2. The hybrid network employs a dual-branch structure, where an ANN branch processes RGB inputs through stacked transformer encoders and an SNN processes event inputs with spiking transformer encoders. Details are provided in Sec. III-B. To enable cross-modal feature interaction, four ISTA adapters are incorporated in each layer of the ANN-SNN encoders. Specifically, two adapters transfer features from ANN to SNN and two from SNN to ANN, with each pair operating at different processing stages. Within the adapter from SNN to ANN, a temporal downsampling attention module aligns multi-step SNN features with single-step ANN features. Finally, the fused features from both branches are added and forwarded to the prediction head to estimate the object center and size in the search region.

### B. ANN and SNN Feature Extraction

*1) ANN Backbone:* For the RGB branch, we utilize vision transformer (ViT) backbone from OSTrack [6]. The RGB template and search images are first transformed into patch embeddings separately, followed by the addition of learnable positional encodings. The resulting tokens are concatenated as $\boldsymbol{x}^I \in \mathbb{R}^{N \times D}$, where $N$ denotes the total number of tokens and $D$ is the embedding dimension. These tokens are subsequently processed by a stack of transformer encoder blocks, each consisting of a multi-head self-attention (MSA) module and a multi-layer perceptron (MLP) module, along
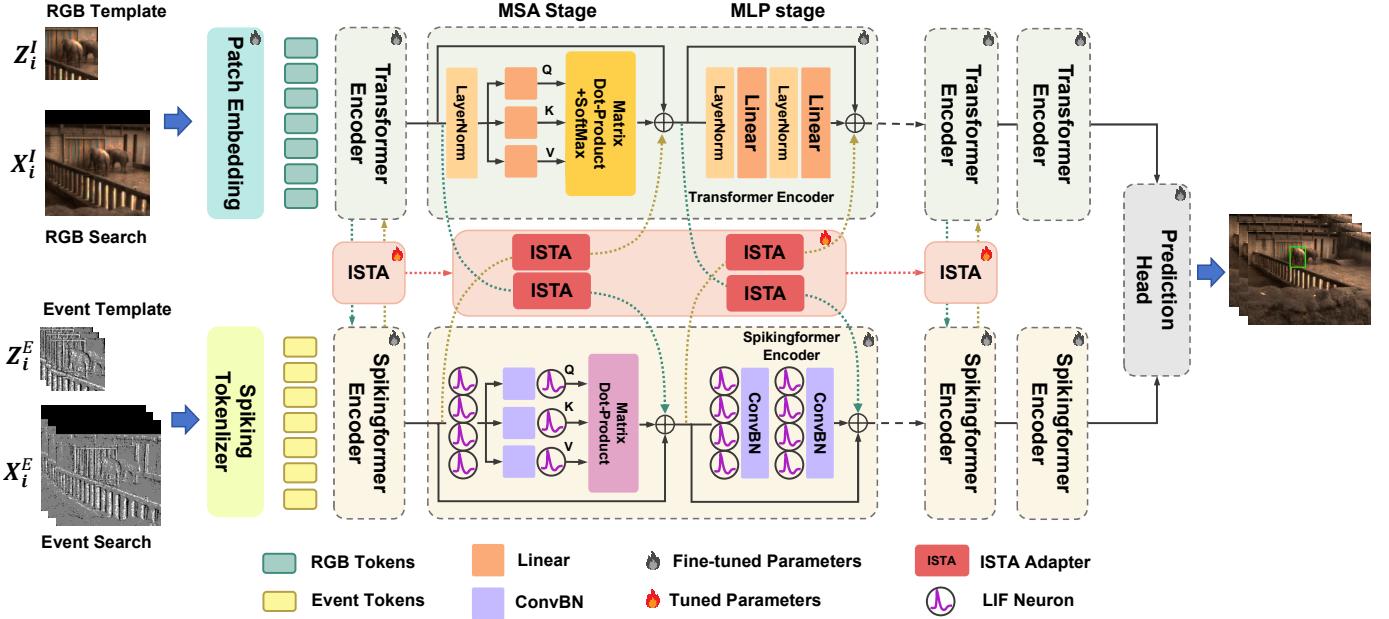
Fig. 2. Overview of the ISTATrack architecture. The hybrid network incorporates a vision transformer branch for RGB inputs and a spiking transformer branch for event data. ISTA adapters are incorporated in the first $K$ transformer encoder layers to enable bidirectional feature adaptation and fusion. During training, ISTA adapters are learned from scratch, while the pretrained backbone parameters are fine-tuned with a smaller learning rate.

with layer normalization and residual connections to enhance training stability. During training, we employ a pretrained model consisting of 12 encoder layers.

*2) SNN Backbone:* In symmetry with the ANN backbone, we apply a modified version of SpikingFormer [64] for the event branch to model global dependencies through spiking self-attention. SpikingFormer is a spike-driven variant of the ViT that significantly reduces energy consumption via the event-driven computing nature of SNNs. It comprises a spiking tokenizer for patch embedding and a series of spiking transformer blocks. It leverages a event-driven residual learning framework based on Leaky Integrate-and-Fire (LIF) neurons, where inputs are first converted into binary spike events (0/1) before linear operations to avoid non-spike computation. Its use of LIF combined with convolutional and batch normalization (ConvBN) layers contributes to energy efficiency. As illustrated in Fig. 2, its encoder differs from the standard ViT architecture. In self-attention computation, event embeddings are processed by LIF neurons and ConvBN layers, followed by another LIF stage to generate spike-form queries (Q), keys (K), and values (V). Attention weights are computed through dot-product operations without softmax due to the non-negativity of spikes. The MLP block is replaced with a lightweight module consisting of two LIF-based ConvBN layers.

The original SpikingFormer is designed for single-image classification. To adapt it for object tracking and ensure architectural symmetry with the ANN backbone, we make two modifications. First, within the spiking tokenizer, the ConvBN layers are shared to process template and search images, whereas separate sets of LIF neurons are used for each to store distinct neuronal states. Second, the original use of 2D convolution in the MLP stage is not suitable for 1D token sequences formed by concatenating template and search

embeddings. We replace them with 1D convolution, allowing joint processing of the combined tokens $\boldsymbol{x}^E \in \mathbb{R}^{N \times D}$.

It is also worth noting that Spikingformer provides a pretrained model with 8 encoder layers, which is inconsistent with the 12-layer ANN backbone. Rather than expanding the spiking encoder to match this depth, we align the 8 spiking layers with the first 8 ANN layers and enable cross-modal feature interaction between them. Results in Sec. IV-E have shown that this design leads to excellent tracking precision while maintaining high efficiency.

*C. ISTA Adapter for Interactive Feature Fusion*

*1) Modeling Cross-Modality Relations based on Sparse Representation:* In the context of sparse representation as introduced in Sec. II-C, a signal can be represented as a linear combination of a small number of atoms from an overcomplete dictionary. For feature embeddings in the transformer encoder, we denote RGB-ANN features as $\boldsymbol{x}^I \in \mathbb{R}^{M \times N}$ and event-SNN features as $\boldsymbol{x}^E \in \mathbb{R}^{T \times M \times N}$, where $M$ is the embedding dimension and $N$ the number of tokens from both template and search. When $T = 1$, the event representation $\boldsymbol{x}^E$ reduces to the same shape as $\boldsymbol{x}^I \in \mathbb{R}^{M \times N}$. Under this condition, we assume that both RGB and event features can be represented using separate dictionaries: $\boldsymbol{x}^I = \boldsymbol{D}_I \boldsymbol{a}^I$ and $\boldsymbol{x}^E = \boldsymbol{D}_E \boldsymbol{a}^E$, where $\boldsymbol{D}_I, \boldsymbol{D}_E \in \mathbb{R}^{M \times D}$ are learned dictionaries and $\boldsymbol{a}^I, \boldsymbol{a}^E \in \mathbb{R}^{D \times N}$ are their corresponding sparse codes. When $T > 1$, the same dictionary is applied across all time steps, and then a temporal downsampling attention module is introduced to aggregate the $T$-step sparse codes into a single-step representation (see Sec. III-D).

To transfer complementary information from the RGB to the event domain, we transform the ANN feature embedding $\boldsymbol{x}^I$ into its SNN counterpart $\boldsymbol{x}^{E'}$. We assume that both feature

embeddings have similar structure and thus share a common sparse code $a^{I \rightarrow E}$, such that we have

$$x^I = D_I a^I,$$
$$x^{E'} = D'_E a^{E'},$$
$$a^I = a^{E'} = a^{I \rightarrow E}.$$
(2)

Similarly, information can be transferred from the event-SNN to the RGB-ANN branch using a shared sparse code $a^{E \rightarrow I}$:

$$x^E = D_E a^E,$$
$$x^{I'} = D'_I a^{I'},$$
$$a^E = a^{I'} = a^{E \rightarrow I}.$$
(3)

These shared sparse codes bridge both modalities and network paradigms, allowing feature projection into a common latent space.

If suitable dictionaries $D_I$ and $D_E$ are available, cross-modal feature adaptation can be formulated as a sparse coding problem:

$$\min_{a^{I \rightarrow E}} \|x^I - D_I a^{I \rightarrow E}\|_2^2 + \lambda \|a^{I \rightarrow E}\|_1,$$
$$\min_{a_i^{E \rightarrow I}} \|x^E - D_E a^{E \rightarrow I}\|_2^2 + \lambda \|a^{E \rightarrow I}\|_1,$$
(4)

where $\|\cdot\|_1$ denotes the $\ell_1$ norm. Once the sparse code $a^{I \rightarrow E^*}$ and $a^{E \rightarrow I^*}$ are obtained, the corresponding transformed feature embeddings can be constructed by

$$x^{E'} = D'_E a^{I \rightarrow E^*},$$
$$x^{I'} = D'_I a^{E \rightarrow I^*}.$$
(5)

In fact, Eq. (4) defines a LASSO problem that can be solved using the iterative shrinkage and thresholding algorithm [27].

*2) ISTA Adapter:* Based on the ISTA algorithm for solving Eq. (4), and omitting modality subscripts for clarity, the sparse coefficients $a$ for a given signal $x$ are iteratively updated following

$$a^k = h_\theta(a^{k-1} + P(x - D a^{k-1})),$$
(6)

where $k = 1, \cdots, K$ denotes the number of iterations, and $h_\theta(x) = sign(x)(x - \theta)_+$ is the soft thresholding function with a threshold $\theta = \boldsymbol{\theta}$. Here $P = D^T$, and we treat $P$ and $D^T$ as independent matrices. The ISTA algorithm iteratively optimizes $a$ with Eq. (6) until convergence criteria are satisfied.

Following the algorithm unfolding strategy [29], we construct a stack of ISTA adapters by unrolling several ISTA iterations into learnable modules. The $k$-th iteration leads to the $k$-th ISTA adapter, as illustrated in Fig. 3(a).

As shown in Fig. 3(b), the operation of an ISTA adapter $\mathcal{A}^{E \rightarrow I}$ from the event-SNN to the RGB-ANN branch can be formulated as

$$a_k = a_{k-1} + P_k^E(x_k^E - D_k^E a_{k-1}),$$
$$a_k = \frac{1}{2}\left(h_{\theta_k}(a_k) + a_k\right),$$
$$a_k = \mathcal{D}_T(a_k) \qquad \text{optional if} \quad T > 1,$$
$$x_k^{I'} = D_k^{I'} a_k,$$
(7)

where the first three equations correspond to the sparse coding stage and the last one represents the feature synthesis stage. In the sparse coding stage, distinct $D_k$, $P_k$, and $\theta_k$ are used at each iteration, where dictionaries $D_k$, $P_k$ are implemented as linear layers and the threshold $\theta_k \in \mathbb{R}^D$ is a learnable vector for each latent element. A skip connection around the soft-thresholding activation is used to promote stable training. When $T > 1$, the TDA module $\mathcal{D}_T$ aggregates multi-step sparse codes into a single-step representation. In the feature synthesis stage, the sparse codes are projected by a synthesis dictionary $D_k^{I'}$ to reconstruct the feature embeddings for the RGB-ANN branch. A similar procedure can be applied to map features from the RGB-ANN branch to the event-SNN branch.

The optimization process is initialized by computing the initial sparse code as $a_0 = P_0 x_0$, where $x_0$ denotes the input feature embedding before entering the transformer encoder and $P_0$ is a linear layer, as depicted in Fig. 3(a). Each ISTA adapter propagates its output sparse code to the subsequent layer, from layer $k$ to $(k+1)$. In this way, ISTA adapters allow iterative refinement of sparse representation and achieve progressive feature adaptation across the transformer encoder layers.

*3) Interactive Feature Adaptation:* An individual ISTA adapter enables feature adaptation from one modality to the other, and thus two adapters are required for bidirectional interaction between RGB-ANN and event-SNN branches. In addition, inspired by [14], we further embed ISTA adapters into both the MSA and MLP stages, resulting in four ISTA adapters for each transformer encoder layer.

As illustrated in Fig. 2 and Fig. 3(c), in the $k$-th transformer encoder layer, the event-SNN branch receives complementary information from the RGB-ANN branch through two ISTA adapters. The feature updates are formulated as

$$x_k^{E^1} = x_k^E + SpikeMSA(x_k^E) + \mathcal{A}_{k1}^{I \rightarrow E}(x^I),$$
$$x_k^{E^2} = x_k^{E^1} + SpikeMLP(x_k^E) + \mathcal{A}_{k2}^{I \rightarrow E}(x^{I^1}),$$
(8)

where $\mathcal{A}_{k1}^{I \rightarrow E}(\cdot)$ and $\mathcal{A}_{k2}^{I \rightarrow E}(\cdot)$ denotes the two ISTA adapters at the MSA and MLP stages at $k$-th encoder. Here, $x_k^I$ and $x_k^{I^1}$ are the input feature embeddings for the two stages in the RGB-ANN branch, respectively, while $x_k^{E^1}$ and $x_k^{E^2}$ denote the corresponding outputs of the SpikeMSA and SpikeMLP stages in the event-SNN branch.

In the reverse direction from event-SNN to RGB-ANN, the interaction is given by

$$x_k^{I^1} = x_k^I + MSA(x_k^I) + \mathcal{A}_{k1}^{E \rightarrow I}(x^E),$$
$$x_k^{I^2} = x_k^{I^1} + MLP(x_k^I) + \mathcal{A}_{k2}^{E \rightarrow I}(x^{E^1}).$$
(9)

This design enables bidirectional and stage-wise feature adaptation between RGB-ANN and event-SNN branches, facilitating dynamic cross-modal interaction and progressive refinement within each transformer encoder layer.

### D. Temporal Downsampling Attention for Time Alignment

In our framework, the SNN branch produces multi-step features, while the ANN branch operates on single-step features. Aligning such temporally asynchronous representation is essential for effective cross-modal fusion. Most existing studies
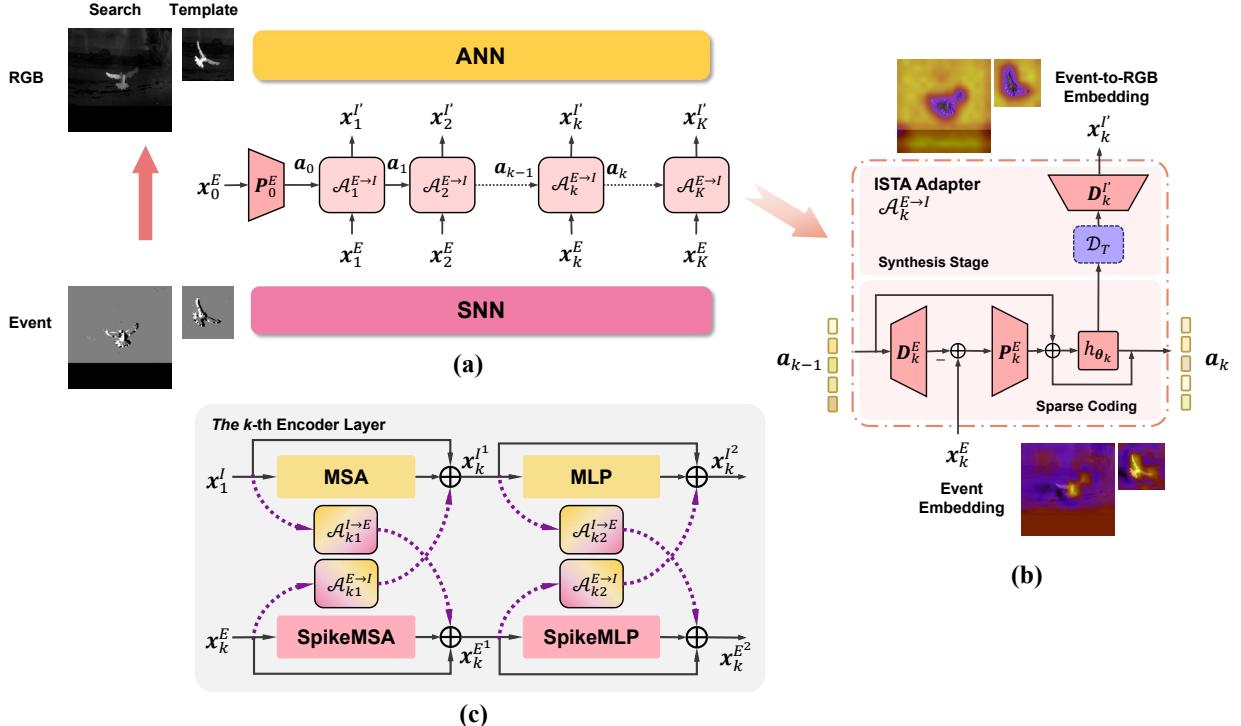
Fig. 3. ISTA adapters for bidirectional feature adaptation and fusion. (a) ISTA adapters $\mathcal{A}^{E \to I}$ from SNN to ANN. The initial sparse codes $\boldsymbol{a}_0$ are generated using the input tokens $\boldsymbol{x}_0^E$. (b) Structure of the $k$-th ISTA adapter, composed of a sparse coding stage and a synthesis stage, where $\mathcal{D}_T$ is an optional TDA module for multi-step SNNs. (c) Four adapters are integrated into the $k$-th transformer encoder layer at both the MSA and MLP stages.

directly average SNN features over time, neglecting the rich temporal dynamics inherent in event-based data. To address this limitation, we propose a temporal downsampling attention (TDA) module that performs adaptive temporal compression while preserving informative time variation.

As depicted in Fig. 4, the TDA module first receives a sequence of SNN features $\boldsymbol{x}^E \in \mathbb{R}^{T \times M \times N}$. Adaptive average pooling and adaptive max pooling are applied separately to the reshaped feature across the temporal dimension, followed by a shared linear operation. Two resulting features are added and then activated by a sigmoid function to produce an attention score vector $\boldsymbol{\alpha} \in \mathbb{R}^{1 \times T}$. This attention score serves as a temporal weighting over the original multi-step feature sequence. The temporally downsampled feature is computed by weighted summation:

$$\boldsymbol{x}_d^E = \boldsymbol{\alpha} \boldsymbol{x}^E \qquad (10)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\boldsymbol{x}_d^E \in \mathbb{R}^{1 \times (M \times N)}$ is finally reshaped into $\mathbb{R}^{M \times N}$ to match the format of single-step ANN features.

The TDA module is embedded within the ISTA adapters from the event-SNN to the RGB-ANN branch, as shown in Eq. (7), aligning the multi-step sparse codes with the single-step ANN representation during feature adaptation. By applying temporal downsampling in the latent sparse space, the TDA module improves memory efficiency while allowing semantic temporal fusion.



Fig. 4. Temporal downsampling attention (TDA) module $\mathcal{D}_T$.

### E. Computational Cost Estimation

An important advantage of incorporating an SNN branch for event data processing is its high energy efficiency. To quantify the efficiency of our ISTASTrack, we estimate the network computational cost following the widely-used approach described in [64]–[66].

In ANNs, computation is dominated by floating-point multiply–and–accumulate (MAC) operations, and the total

number of floating-point operations (FLOPs) is typically used to approximate the computational workload. In contrast, SNNs leverage the event-driven computing nature, which significantly reduces the computational cost. In SNNs, synaptic operations (SyOPs) refer to the accumulation of postsynaptic inputs triggered by presynaptic spike events. For linear layers following spiking activities, such as convolutional or fully connected layers, these operations reduce to accumulation (AC) only, since multiplication can be removed owing to the binary spikes (0/1). The number of SyOPs is therefore estimated as $fr \times OP_{AC}$, where $fr$ denotes the average neuronal firing rate and $OP_{AC}$ the peak number of accumulation operations in the layer. Nevertheless, SNNs also require certain non-synaptic computations that still rely on MAC operations, similar to ANNs. Consequently, the computational energy consumption of our hybrid network can be estimated by

$$E = E_{MAC} \times \sum_i OP^i_{MAC} + E_{AC} \times \sum_i (OP^i_{AC} \times fr^i) \quad (11)$$

where $i = 1 \times N$ indexes the network layers, $E_{MAC} = 4.6pJ$ and $E_{AC} = 0.9pJ$ represent the approximate energy cost of 32-bit floating-point MAC and AC operations on 45nm hardware, respectively [67]. A detailed analysis of computational efficiency can be found in Sec. IV-F.

## IV. EXPERIMENTAL RESULTS

### A. Experimental Settings

### B. Datasets

We utilize four representative RGB-Event tracking datasets, including FE240hz [9], COESOT [63], VisEvent [10], and FELT [40]. The four datasets are all collected by DAVIS346, an RGB-event hybrid camera, with an image resolution of $346 \times 260$.

The FE240hz dataset covers various real-world challenges such as low lighting, high dynamic range, motion blur, and fast object motion. However, the camera is stationary in most scenes, where the event frames exhibit relatively clean backgrounds. It provides high-frequency (240Hz) annotations, but we only use those synchronized with the RGB frame rate. The dataset consists of 140K annotated frames, with 101 sequences used for training and 40 for testing.

The VisEvent dataset focuses primarily on traffic scenes and includes targets such as pedestrians and vehicles, along with a few indoor scenarios. It comprises approximately 370K frames across 820 paired RGB and event sequences. Due to the incomplete raw event data, 295 sequences are used for training and 219 for testing in our setting.

The COESOT dataset consists of 1,354 aligned sequences, with 827 sequences for training and 527 for testing. Compared to other datasets, COESOT provides a broader range of target categories, including vehicles, pedestrians, and animals such as birds, monkeys, and elephants.

The FELT dataset is the largest available benchmark for long-term RGB-Event VOT, comprising 742 sequences and approximately 1.6M paired RGB frames and event streams. It features extended sequences with more complex scenarios, including various ball sports, with very small objects and

significant background interference. The dataset is divided into 520 training sequences and 222 testing sequences.

*1) Training Details:* To train our network, we employ a composite loss function comprising focal loss for classification and a combination of L1 loss and GIoU loss for bounding box regression:

$$L = \lambda_1 L_{\text{focal}} + \lambda_2 L_1 + \lambda_3 L_{\text{GIoU}}, \quad (12)$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are trade-off coefficients, empirically set to 2, 5, and 1, respectively, following the setting in [6].

For initialization, we load the pretrained weights from OS-Track [6] for the ANN branch and partially initialize the SNN branch using a pretrained SpikingFormer checkpoint [64], excluding the modified layers. During training, the pretrained parameters are fine-tuned with a smaller learning rate of $1e^{-5}$, while the remaining parameters are trained with a learning rate of $1e^{-4}$. The network is trained for 60 epochs using the Adam optimizer with a weight decay of $1e^{-4}$ and a batch size of 32. The learning rate is reduced to 10% of its initial value after 48 epochs. Input image sizes are set to $128 \times 128$ for the template and $256 \times 256$ for the search region. The number of SNN time steps $T$ is set to 3 by default to achieve a balance between accuracy and computational efficiency. The number of layers integrating ISTA adapters, denoted as $N$, is set to 4 by default, corresponding to the configuration that yields optimal performance (see Sec. IV-E).

*2) Evaluation Metrics:* We evaluate tracking performance using several metrics, including success rate (SR), overlap precision ($OP_T$), precision (PR), and normalized precision (NPR). The SR measures the proportion of frames where the intersection-over-union (IoU) between the predicted and ground truth bounding boxes exceeds a range of thresholds, and we report the area under the curve (AUC) of the SR plot as the final score. $OP_T$ refers to the SR at a specific IoU threshold T; we report results for OP50 and OP75, corresponding to thresholds of 50% and 75%, respectively. PR measures the percentage of frames where the Euclidean distance between the predicted and ground truth bounding box centers falls below a given threshold, which we set to 20 pixels. NPR follows the same definition as PR but normalizes the centers by the ground truth bounding box dimensions. Frames where the target is out of view are excluded from evaluation.

### C. Comparison with State-of-the-art Trackers

*1) Quantitative Comparison:* We compare our method against a wide range of state-of-the-art trackers, covering three main categories: single-branch ANN, dual-branch ANN, and hybrid ANN-SNN approaches. Results are summarized in Tab. I. For fair comparison, we retrain nearly all networks on the four aforementioned datasets using our settings. For CrossEI and SNNPTrack, where code or checkpoints are not publicly available, we report the results as published in their respective papers.

**Single-branch ANN**. Classic ANN-based trackers originally designed for single-modality tasks can be adapted for RGB-Event tracking by modifying their first layer to process early-fused inputs, including discriminative learning-based

TABLE I
COMPARISON OF TRACKING RESULTS WITH OTHER STATE-OF-THE-ART TRACKERS. TRACKERS MARKED WITH * DISPLAY SCORES PUBLISHED IN PRIOR RESPECTIVE PAPERS. **BOLD** VALUES REPRESENT THE BEST RESULTS FOR THE SAME DATASET.

| Type | Tracker | Architecture | FE240hz | | | | | VisEvent | | | | | COESOT | | | | | FELT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | SR | OP50 | OP75 | PR | NPR | SR | OP50 | OP75 | PR | NPR | SR | OP50 | OP75 | PR | NPR | SR | OP50 | OP75 | PR | NPR |
| Single-branch ANN | ATOM [32] | ResNet50 | 38.7 | 49.0 | 19.3 | 57.1 | 43.8 | 46.4 | 55.0 | 30.7 | 64.7 | 54.6 | 59.0 | 70.5 | 51.7 | 70.4 | 70.0 | 40.3 | 46.5 | 33.5 | 47.9 | 44.7 |
| | DiMP [5] | ResNet50 | 60.2 | 77.4 | 30.3 | 90.5 | 69.2 | 50.5 | 60.3 | 35.5 | 70.1 | 58.8 | 62.6 | 74.4 | 54.0 | 73.9 | 73.7 | 48.7 | 57.7 | 40.9 | 59.1 | 55.0 |
| | PrDiMP [33] | ResNet50 | 60.6 | 77.6 | 30.0 | 91.7 | 68.7 | 50.7 | 59.8 | 38.0 | 69.1 | 59.0 | 61.6 | 72.2 | 55.0 | 71.3 | 71.4 | 49.4 | 57.6 | 42.4 | 59.7 | 55.2 |
| | STARK [35] | ResNet+ViT | 43.1 | 54.0 | 20.0 | 65.5 | 46.8 | 37.5 | 41.8 | 27.0 | 49.2 | 41.6 | 61.9 | 70.9 | 56.0 | 70.3 | 69.9 | 54.3 | 61.0 | 45.6 | 70.9 | 58.1 |
| | OSTrack [6] | ViT | 58.6 | 72.5 | 32.3 | 88.8 | 64.7 | 59.6 | 72.3 | 51.9 | 78.2 | 71.4 | 65.1 | 74.3 | 61.1 | 73.5 | 73.2 | 50.9 | 58.8 | 44.8 | 61.3 | 56.5 |
| | AiATrack [37] | ViT | 55.8 | 70.7 | 25.6 | 85.9 | 60.3 | 49.1 | 58.9 | 35.5 | 67.0 | 57.4 | 65.1 | 76.3 | 58.1 | 75.4 | 74.2 | 54.8 | 61.9 | 46.7 | **69.4** | 58.9 |
| | CEUTrack [63] | Unified ViT | 43.5 | 52.4 | 21.0 | 66.4 | 48.2 | 59.2 | 72.9 | 50.2 | 77.8 | 71.1 | 67.5 | 78.6 | 65.7 | 78.6 | 77.7 | 50.9 | 59.3 | 46.2 | 61.2 | 57.1 |
| Dual-branch ANN | FENet [8] | Dual ResNet18 | 52.0 | 63.7 | 21.2 | 81.1 | 57.1 | 44.4 | 52.4 | 34.9 | 60.0 | 51.0 | 53.2 | 59.2 | 41.5 | 56.1 | 57.4 | 47.6 | 55.0 | 42.2 | 57.3 | 52.9 |
| | AFNet [9] | Dual ResNet18 | 57.8 | 73.1 | 28.9 | 87.8 | 63.7 | 41.7 | 48.1 | 32.9 | 54.8 | 47.2 | 53.8 | 58.6 | 43.5 | 57.1 | 58.7 | 40.4 | 44.4 | 30.0 | 46.7 | 42.7 |
| | CrossEI* [11] | Dual ResNet18 | 59.4 | 74.6 | 28.8 | **92.3** | - | 53.1 | 62.4 | 40.3 | 71.4 | - | 61.7 | 70.1 | 59.0 | 70.9 | - | - | - | - | - | - |
| | TENet [42] | ViT+Pooler | 60.3 | 76.9 | 35.0 | 88.4 | 67.7 | 65.1 | 78.1 | 61.9 | 82.1 | 77.2 | 74.7 | 85.7 | 75.2 | 85.7 | 85.0 | 54.3 | 62.2 | 50.8 | 64.3 | 59.9 |
| | ViPT [43] | Dual ViT, Vision Prompt | 61.0 | 78.1 | 36.8 | 88.3 | 69.1 | 64.2 | 77.1 | 60.1 | 81.4 | 75.9 | 73.6 | 84.2 | 73.9 | 84.3 | 83.3 | 51.6 | 59.1 | 47.4 | 61.1 | 56.8 |
| | UnTrack [13] | | 62.0 | 79.4 | 36.4 | 90.1 | 69.2 | 62.0 | 73.5 | 56.1 | 78.8 | 72.4 | 75.3 | 85.4 | 75.9 | 85.5 | 84.6 | 52.6 | 60.2 | 48.8 | 61.9 | 57.9 |
| | BAT [45] | | 64.3 | 82.3 | 41.6 | 91.9 | **73.1** | 67.1 | 81.0 | **63.7** | **84.8** | 79.4 | 75.6 | 85.3 | 76.7 | 85.7 | 84.4 | 52.9 | 60.3 | 49.1 | 62.6 | 58.0 |
| ANN-SNN | MMHT [20] | ResNet18+ AlexSNN | 47.6 | 60.1 | 16.8 | 73.8 | 50.9 | 43.7 | 53.6 | 22.8 | 63.6 | 51.9 | 60.2 | 72.2 | 50.0 | 68.5 | 70.3 | 28.4 | 32.1 | 10.0 | 35.3 | 30.4 |
| | SNNPTrack* [21] | ViT+SNN Prompt | - | - | - | - | - | 59.8 | - | - | 76.9 | - | 66.8 | - | - | 74.8 | - | - | - | - | - | - |
| | ISTASTrack | ViT+ SNN ViT | **64.7** | **82.3** | **43.3** | **92.2** | **73.1** | **67.3** | **81.2** | 63.5 | 84.6 | **79.7** | **75.7** | **86.8** | **76.9** | **87.1** | **86.0** | **55.2** | **63.1** | **51.4** | 65.8 | **60.8** |

models such as ATOM [32], DiMP [5], and PrDiMP [33], as well as transformer-based designs like STARK [35], OSTrack [6], and AiATrack [37]. CEUTrack [63] further introduces an early fusion module for ViT, specifically tailored to RGB-Event tracking. As shown in Tab. I, early-fusion ANN trackers remain limited in effectiveness, in which transformer-based models such as OSTrack, AiATrack, and CEUTrack achieving relatively stronger performance.

**Dual-branch ANN.** This category includes methods explicitly designed for RGB-Event fusion, such as FENet [8], AFNet [9], and CrossEI [11], which adopt dual ResNet-based backbones. Their effectiveness is constrained by relatively simple feature extractors and complex inference pipelines, though CrossEI achieves better results with its motion-oriented design. In contrast, TENet [42] incorporates a ViT backbone for the RGB branch while tailoring the event branch to event-specific characteristics, delivering competitive performance. We also consider dual-modality transformer models based on VPT, including ViPT [43], UnTrack [13], and BAT [45]. These VPT-based frameworks substantially outperform most models, except for TENet and our ISTASTrack, among which BAT achieves the second-best overall results. This highlights the importance of employing dual-branch architectures with well-designed fusion modules or prompts for processing different modalities, which often surpass early-fusion approaches in handling multimodal inputs.

**ANN-SNN.** As introduced in Sec. I, only two hybrid ANN-SNN models have been proposed for RGB-Event tracking, i.e. MMHT [20] and SNNPTrack [21]. MMHT utilizes simple ResNet and AlexSNN as backbones, resulting in relatively weak performance. By contrast, SNNPTrack employs SNN-based prompts inserted between ViT encoders to guide fusion, demonstrating stronger results. However, its unidirectional design limits the ability to fully exploit advantages from both

modalities. In comparison, our ISTASTrack employs an SNN-ViT backbone for the event branch and integrates bidirectional adapters for enhanced RGB–Event fusion, thereby fully leveraging complementary spatio-temporal information and achieving the best overall performance. Results of ISTASTrack are the best across different SNN time steps, details can be found in Tab. VI.

*2) Visual Comparison of Typical Scenes:* To further illustrate the comparison, Fig. 5 presents tracking results of representative methods across typical scenarios, including low light, overexposure, small objects, similar objects, fast motion, and occlusions. In the visualization, green boxes denote the ground truth (GT) target, while red boxes indicate the predictions of our ISTASTrack. The results demonstrate that our network remains robust under challenging lighting, fast motion, and occlusions, and can detect small objects more accurately while distinguishing them from similar distractors.

Furthermore, Fig. 6 presents two sequences involving similar objects and occlusions. In the left example, showing a football game, ISTASTrack successfully tracks the target player despite occlusions from the goalpost and other players, whereas other methods lose the target and often switch to a different person. Similarly, in the right example with multiple flying birds, our method reliably re-identifies the correct target after they intermingle, demonstrating strong robustness in complex scenarios.

*3) Attribute-based Comparison:* Since VisEvent, COESOT, and FELT datasets provide sequence-level attributes, we plot a radar chart in Fig. 7 to compare ISTASTrack with representative methods across different scene attributes. Overall, ISTASTrack outperforms other methods on nearly all attributes, particularly on VisEvent. Transformer-based methods with interactive fusion, such as TENet and UnTrack, also show stronger robustness than the remaining baselines. For
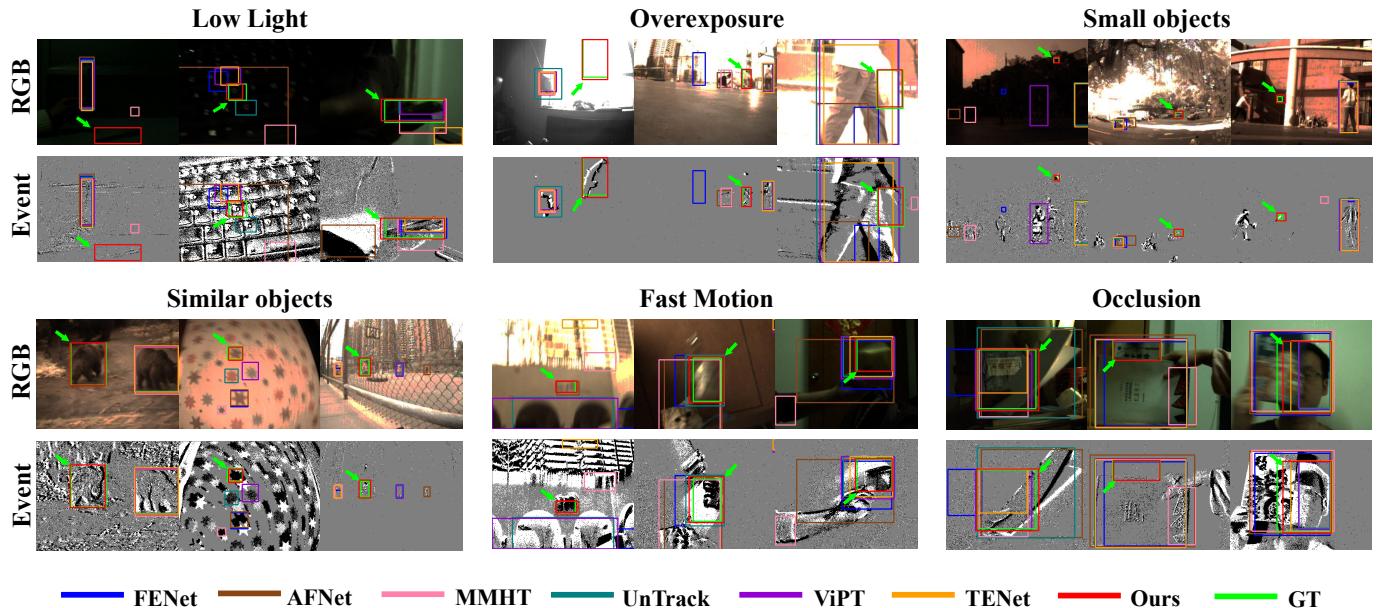
Fig. 5. Qualitative comparison of tracking results across challenging scenarios, including low light, overexposure, small objects, similar objects, fast motion, and occlusions. Green boxes and arrows indicate ground truth (GT) targets, and red boxes represent predictions from ISTASTrack, which demonstrates strong robustness and accurate target discrimination under diverse conditions.
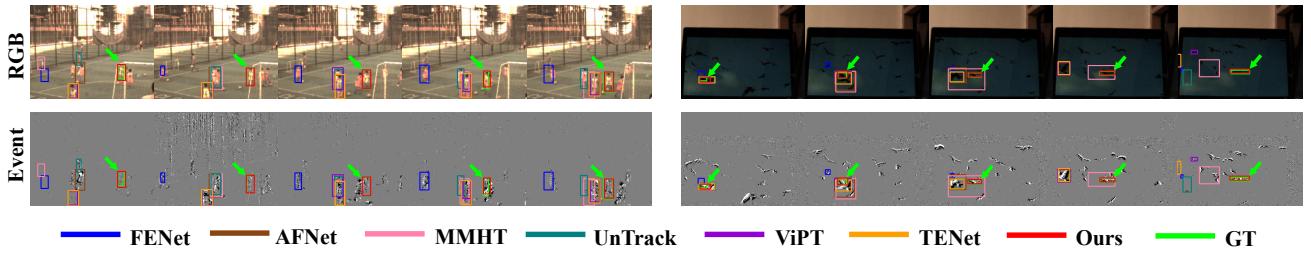


Fig. 6. Tracking results on challenging long sequences with similar objects and occlusions. Green boxes and arrows indicate ground truth (GT) targets, and red boxes represent predictions from ISTASTrack. Our method consistently maintains accurate target tracking despite occlusions and object confusion.

challenging conditions like partial occlusion (POC) and full occlusion (FOC), all models see a drop in performance, yet ISTASTrack demonstrates greater resilience to occlusion. On the most challenging FELT dataset, where fast motion (FM), strong background interference (BI), and FOC lead to generally low performance, our method performs slightly better than the other approaches despite the overall difficulty.

### D. Visualization of Intermediate Features

To demonstrate the effectiveness of the ISTA adapter in multimodal adaptation, we visualize intermediate features from the first encoder layer in Fig. 8. Specifically, we present the average values of feature maps over time for the search area, showing ISTA adapter outputs from RGB to event (i.e., RGB2E feature) and from event to RGB (i.e., E2RGB feature), along with the corresponding ANN and SNN attention maps (i.e., Attn / adapter). For comparison, we also provide ANN and SNN attention maps from a model without the ISTA adapter (i.e., Attn w/o adapter), with these baseline results marked by red bounding boxes.

We select four examples that feature challenging lighting, background distractions, and similar objects. In the first row,

severe overexposure and underexposure make the target difficult to identify in the RGB image, causing the ANN attention maps without ISTA adapters to miss the unseen target. In the model with ISTA adapters, the ANN successfully locks onto the target, while the SNN also achieves more accurate focus by leveraging ISTA features. In the last row, where scenes contain rich textures, the ANN can accurately localize the target, but the SNN attention becomes dispersed due to similar contours and the absence of color cues. In contrast, features from the ISTA adapter guide the SNN to concentrate on the correct target.

### E. Ablation Study

*1) Single Modality vs. Dual Modality:* Tab. II compares the performance of single- and dual-modality models. "RGB-ANN" and "Event-SNN" denote models trained only on RGB and event data via the ANN and SNN branches, respectively. Our hybrid ISTASTrack, which uses both modalities, outperforms single-modality models across all four datasets. In particular, Event-SNN surpasses RGB-ANN on the FE240hz dataset, which features extreme lighting conditions that favor event data. In contrast, on other three datasets, RGB-ANN
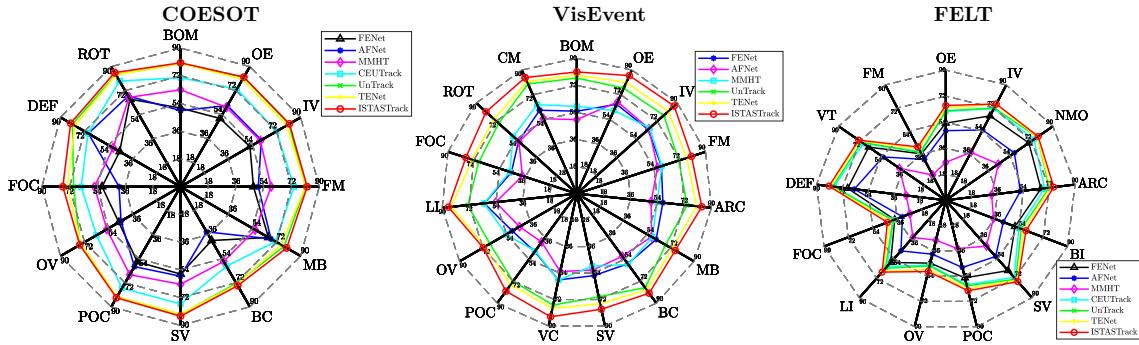
Fig. 7. Precision scores comparison of different attributes on COESOT, VisEvent and FELT datasets.
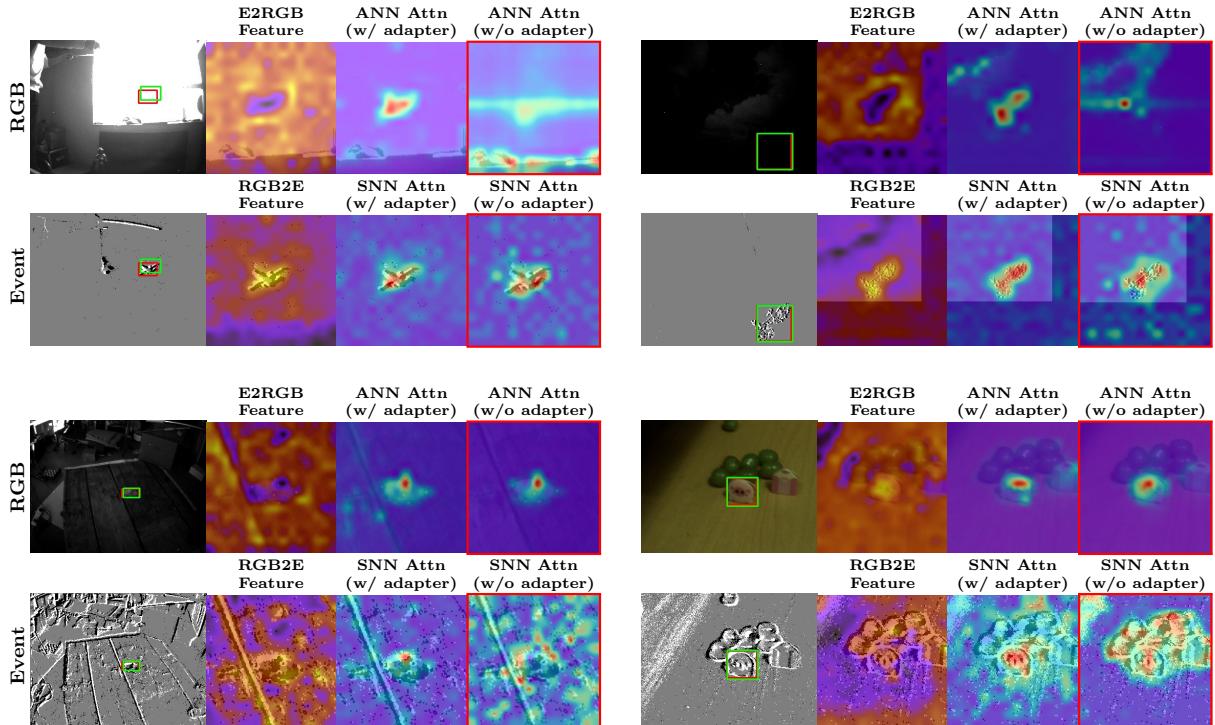


Fig. 8. Visualization of ISTA features and attention maps from the first encoder layer. RGB2E and E2RGB represent feature maps transferred by the ISTA adapter from RGB to event and from event to RGB, respectively. ANN and SNN attention maps are shown (i.e., Attn w/ adapter) alongside their counterparts from a model without the ISTA adapter (i.e., Attn w/o adapter, highlighted with red bounding boxes) for comparison.

TABLE II
COMPARISON OF TRACKING RESULTS OF SINGLE- AND DUAL-MODALITY MODELS. **BOLD** VALUES REPRESENT THE BEST RESULTS FOR THE SAME DATASET.

| Model | FE240hz | | VisEvent | | COESOT | | FELT | |
|---|---|---|---|---|---|---|---|---|
| | SR | PR | SR | PR | SR | PR | SR | PR |
| **RGB-ANN** | 50.85 | 73.49 | 64.74 | 81.38 | 74.41 | 84.82 | 52.36 | 61.03 |
| **Event-SNN** | 53.84 | 80.71 | 38.93 | 52.58 | 52.80 | 59.10 | 38.51 | 44.35 |
| **Hybrid** | **64.59** | **92.83** | **67.26** | **84.61** | **75.74** | **86.82** | **55.18** | **65.75** |

significantly outperforms Event-SNN and nearly matches the hybrid model, indicating a stronger reliance on the RGB modality.

*2) Effectiveness of SNN Backbone:* To evaluate the effectiveness of the SNN branch, we conduct an experiment re-

placing the ANN-SNN architecture with dual ANN branches. As shown in Tab. III, the SNN backbone achieves comparable results to the dual-ANN setup, even performing slightly better on the FE240hz and VisEvent datasets. Notably, the SNN branch has a shallower depth of 8 compared to that of 12 for the ANN branch, and it offers significantly higher energy efficiency, which will be further analyzed in Sec. IV-F.

*3) Effectiveness of ISTA Adapter and TDA Module:* In addition to the qualitative demonstration of the effectiveness of the ISTA adapter in Fig. 8, we further perform an ablation study to quantitatively evaluate different ISTA configurations and the function of the TDA module, as summarized in Tab. IV.

Compared to the baseline model without ISTA adapters (the first row), incorporating unidirectional adapters (RGB-

TABLE III
COMPARISON OF TRACKING RESULTS OF NETWORKS WITH DUAL-ANN AND ANN-SNN BACKBONES. **BOLD** VALUES REPRESENT THE BEST RESULTS FOR THE SAME DATASET.

| Model | FE240hz | | VisEvent | | COESOT | | FELT | |
|---|---|---|---|---|---|---|---|---|
| | SR | PR | SR | PR | SR | PR | SR | PR |
| ANN-ANN | 64.02 | 92.56 | 66.68 | 83.33 | 75.73 | **87.03** | **55.31** | **65.85** |
| ANN-SNN | **64.59** | **92.83** | **67.26** | **84.61** | **75.74** | 86.82 | 55.18 | 65.75 |

to-Event or Event-to-RGB) results in noticeable performance improvements except for the FELT dataset, with the Event-to-RGB adapter providing slightly greater gains. The model equipped with bidirectional adapters enables information exchange in both directions, consistently achieving the highest SR and PR scores. This highlights the benefit of mutual feature adaptation for effective multimodal fusion. Furthermore, removing the TDA module (w/o TDA) leads to lower tracking accuracy, indicating its crucial role in aggregating semantic temporal information from multi-step event features.

*4) Adapter Number and Location:* Since ANN and SNN branches have different transformer encoder depths (12 for ANN and 8 for SNN), both the number and placement of ISTA adapters influence the fusion performance. As shown in Tab. V, inserting adapters in the first $N$ layers (the first four rows) generally yields higher SR and PR scores than placing them in the last $N$ layers (the last two rows), suggesting that early feature interaction is more effective than late fusion.

In terms of quantity, adding more adapters does not necessarily yield better results, with four layers of adapters achieving the best overall performance. Once early cross-modal information is exchanged, each branch can effectively process the integrated features independently, whereas additional adapters in deeper layers may interfere with these refined representations. Compared to the model without adapters (see Tab. IV), incorporating only two layers of ISTA adapters already brings substantial improvements, demonstrating effective cross-modal feature adaptation and fusion between ANN and SNN branches.

*5) Number of Time Steps:* Increasing the number of time steps of event data generally preserves richer temporal cues and yields informative temporal feature representations. We investigate this effect by varying the number of time steps $T$ for both the single-modality SNN and the hybrid model.

As shown in Tab. VI, "Event-SNN" refers to SNN models that take event data as inputs, while "Hybrid" denotes the ANN-SNN model with RGB–Event inputs. Across the four datasets, setting $T = 3$ significantly improves SR and PR for Event-SNN compared to $T = 1$. However, this trend is not consistently observed in the hybrid model, except for the FE240hz dataset. A plausible reason is that on VisEvent, COESOT, and FELT datasets, the RGB modality already achieves strong performance, while their event data are relatively weak. In these cases, tracking accuracy of Event-SNN approaches that of the hybrid model, as reported in Tab. II. In contrast, the FE240hz dataset provides more balanced modalities, where using the event input alone can outperform the model using the RGB input, and the hybrid model achieves significantly

better performance than either modality alone. Consequently, improvements of Event-SNN do not necessarily lead to higher hybrid performance on VisEvent, COESOT, and FELT, but they do on FE240hz.

We further evaluate more time steps ($T = 5$ and $T = 7$) on FE240hz. As shown in Tab. VI, increasing $T$ enhances the tracking performance for both Event-SNN and hybrid models. In the hybrid case, the gains are particularly evident in OP75, reflecting more precise localization and size estimation of moving objects. For instance, increasing $T$ from 1 to 5 improves OP75 from 39.84% to 43.27%. However, the benefit becomes marginal with larger $T$, and the hybrid model even shows a slight drop at $T = 7$. Moreover, more time steps impose higher computational and memory costs, highlighting a trade-off between accuracy and efficiency. In practice, we find that $T = 3$ already offers competitive results while effectively leveraging the advantages of multi-step SNNs.

### F. Computational Efficiency

Tab. VII presents the computational cost of ISTASTrack with different event backbone architectures. The ANN–ANN configuration has more parameters than the ANN–SNN variant, since the latter employs a shallower SNN branch. Computational consumption of the ANN-ANN model results from MAC operations (56.4G), resulting in a total of 259mJ. In contrast, the hybrid ANN–SNN design reduces MACs to about 60% of the ANN baseline, with a slight increase as $T$ grows due to additional non-synaptic operations in the SNN. In terms of FLOPs, the ANN branch with ISTA adapters and the prediction head contributes the majority, while the SNN branch accounts for under 10% when $T = 1$. ACs and SyOps are significantly influenced by the number of time steps $T$, and their energy consumption is proportional to $T$. Nonetheless, most of the energy consumption originates from non-spiking computations. Increasing $T$ does not significantly increase the overall computational cost. These results demonstrate that the hybrid architecture achieves higher energy efficiency by leveraging sparse spike events and low-energy ACs in the SNN branch.

In addition, we report the metrics of the ISTA adapter and the TDA module for the model with $T = 3$. ISTA adapters have only 0.32M parameters, which is approximately 2% of the total, thus incurring only a marginal energy cost. The TDA module is even more lightweight and its contribution to computational cost is practically negligible. These results demonstrate that our proposed modules are very efficient for cross-modal and cross-paradigm fusion.

## V. CONCLUSION

In this work, we propose ISTASTrack, a hybrid ANN–SNN tracker with ISTA adapters for RGB-Event VOT. The hybrid network combines a vision transformer for RGB inputs and a spiking transformer for event streams, leveraging the complementary strengths of the two modalities. Between transformer encoders in the ANN and SNN branches, we design ISTA adapters based on sparse representation theory and algorithm unfolding. These lightweight and interpretable ISTA adapters

TABLE IV
COMPARISON OF TRACKING RESULTS OF ISTASTRACK VARIANTS WITH DIFFERENT ISTA ADAPTER AND TDA CONFIGURATIONS. **BOLD** VALUES
REPRESENT THE BEST RESULTS FOR THE SAME DATASET.

| Model | FE240hz | | VisEvent | | COESOT | | FELT | |
|---|---|---|---|---|---|---|---|---|
| | SR | PR | SR | PR | SR | PR | SR | PR |
| **w/o adapter** | 60.04 | 87.53 | 64.94 | 81.88 | 74.7 | 85.54 | 52.83 | 61.76 |
| **RGB-to-Event adapter** | 61.72 | 90.52 | 65.83 | 83.27 | 75.4 | 86.63 | 52.69 | 61.44 |
| **Event-to-RGB adapter** | 64.29 | **93.16** | 66.16 | 83.67 | 75.53 | 86.94 | 52.21 | 61.28 |
| **w/o TDA** | 63.46 | 91.51 | 66.28 | 83.48 | 73.96 | 84.78 | 53.72 | 63.14 |
| **Bidirectional adapter** | **64.59** | 92.83 | **66.46** | **84.34** | **75.74** | 86.82 | **54.93** | **65.75** |

TABLE V
COMPARISON OF TRACKING RESULTS OF ISTASTRACK VARIANTS WITH
DIFFERENT ADAPTER NUMBERS AND LAYERS. **BOLD** VALUES REPRESENT
THE BEST RESULTS FOR THE SAME DATASET.

| Model | FE240hz | | VisEvent | | COESOT | | FELT | |
|---|---|---|---|---|---|---|---|---|
| (Number, layer) | SR | PR | SR | PR | SR | PR | SR | PR |
| N=2, 1-2 | 63.01 | 90.49 | 65.61 | 83.06 | 75.54 | 86.69 | 54.57 | 64.87 |
| N=4, 1-4 | **64.59** | **92.83** | **66.46** | **84.34** | **75.74** | **86.82** | **54.93** | **65.75** |
| N=6, 1-6 | 63.46 | 91.14 | 65.2 | 82.44 | 74.96 | 86.3 | 54.51 | 64.54 |
| N=8, 1-8 | 63.87 | 92.63 | 65.87 | 83.47 | 74.79 | 86.1 | 54.59 | 64.92 |
| N=4, 8-12 | 63.27 | 91.73 | 65.57 | 82.58 | 75.24 | 86.26 | 54.29 | 64.26 |
| N=8, 4-12 | 62.99 | 91.2 | 65.25 | 82.47 | 75.01 | 85.71 | 54.67 | 65.25 |

TABLE VI
COMPARISON OF TRACKING RESULTS FOR VARYING THE NUMBER OF
TIME STEPS. "EVENT-SNN" REFERS TO SNN MODELS THAT USE EVENT
DATA AS INPUTS, WHEREAS "HYBRID" DENOTES HYBRID NETWORKS
THAT TAKE RGB-EVENT INPUTS. **BOLD** VALUES REPRESENT THE BEST
RESULTS FOR THE SAME DATASET.

| | Time step | Event-SNN | | Hybrid | |
|---|---|---|---|---|---|
| | | SR | PR | SR | PR |
| VisEvent | T=1 | 36.27 | 49.94 | **67.26** | **84.61** |
| | T=3 | **38.93** | **52.58** | 66.46 | 84.34 |
| COESOT | T=1 | 47.38 | 49.82 | 75.63 | 86.8 |
| | T=3 | **52.80** | **59.10** | **75.74** | **86.82** |
| FELT | T=1 | 32.81 | 34.36 | **55.18** | **65.75** |
| | T=3 | **38.51** | **44.35** | 54.94 | **65.75** |
| | | SR | OP50 | OP75 | PR |
| FE240hz (Event-SNN) | T=1 | 51.99 | 65.22 | 23.71 | 79.84 |
| | T=3 | 53.84 | 68.62 | **27.79** | 80.71 |
| | T=5 | 55.44 | 68.95 | 27.52 | 85.8 |
| | T=7 | **56.27** | **69.2** | 27.55 | **87.96** |
| FE240hz (Hybrid) | T=1 | 63.82 | 81.72 | 39.84 | 91.8 |
| | T=3 | 64.59 | 82.28 | 41.61 | **92.83** |
| | T=5 | **64.74** | **82.34** | **43.27** | 92.15 |
| | T=7 | 63.91 | 81.74 | 41.55 | 91.6 |

effectively achieve bidirectional feature interaction and adaptation. Finally, a temporal downsampling attention (TDA) module ensures semantic temporal alignment between multi-step SNN features and single-step ANN features. Experimental results demonstrate that ISTASTrack achieves high tracking accuracy with improved energy efficiency across multiple datasets and challenging scenarios. Furthermore, our comprehensive analyses of the hybrid architecture and feature adaptation strategies provide insightful guidance into the design of hybrid networks and the effective integration of ANN and

SNN modalities.

The main limitation of the proposed architecture is that the potential of SNNs for processing long sequences has not been fully explored due to memory constraints. The current design resets SNN states after each inference step, restricting the ability of the network to leverage long-term temporal dependencies and historical information. Future work will focus on enabling SNNs to adapt and update temporal context across longer sequences, fully leveraging their capability in dynamic temporal modeling for high-speed vision tasks.

## REFERENCES

[1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[2] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4947–4954, Jul 2021.

[3] H. Liu, D. P. Moeys, G. Das, D. Neil, S.-C. Liu, and T. Delbrück, "Combined frame- and event-based detection and tracking," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 2511–2514.

[4] X. Wang, Z. Wu, B. Jiang, Z. Bao, L. Zhu, G. Li, Y. Wang, and Y. Tian, "HARDVS: Revisiting Human Activity Recognition with Dynamic Vision Sensors," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 6, pp. 5615–5623, Mar 2024.

[5] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning Discriminative Model Prediction for Tracking," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct 2019, pp. 6181–6190.

[6] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer Nature Switzerland, 2022, vol. 13682, pp. 341–357.

[7] C. Tang, X. Wang, J. Huang, B. Jiang, L. Zhu, J. Zhang, Y. Wang, and Y. Tian, "Revisiting Color-Event based Tracking: A Unified Network, Dataset, and Metric," Jan 2024. [Online]. Available: http://arxiv.org/abs/2211.11010

[8] J. Zhang, X. Yang, Y. Fu, X. Wei, B. Yin, and B. Dong, "Object Tracking by Jointly Exploiting Frame and Event Domain," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021, pp. 13 023–13 032.

[9] J. Zhang, Y. Wang, W. Liu, M. Li, J. Bai, B. Yin, and X. Yang, "Frame-Event Alignment and Fusion Network for High Frame Rate Tracking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2023, pp. 9781–9790.

[10] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, "VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows," *IEEE Trans. Cybern.*, vol. 54, no. 3, pp. 1997–2010, Mar 2024.

[11] Z. Chen, J. Wu, W. Dong, L. Li, and G. Shi, "CrossEI: Boosting Motion-Oriented Object Tracking With an Event Camera," *IEEE Trans. Image Process.*, vol. 34, pp. 73–84, 2025.

[12] C. Liu, Z. Guan, S. Lai, Y. Liu, H. Lu, and D. Wang, "EMTrack: Efficient Multimodal Object Tracking," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2024.

TABLE VII
COMPARISON OF COMPUTATIONAL COSTS ACROSS DIFFERENT EVENT BACKBONE ARCHITECTURES AND OUR PROPOSED MODULE. "ANN-SNN"
DENOTES THE HYBRID ISTASTRACK, AND "ANN–ANN" DENOTES USING AN ANN IN PLACE OF THE SNN BRANCH.

| Model | Params | MAC(G) | AC(G) | FLOPs(G) ANN | FLOPs(G) SNN | SyOps(G) | $E_{ANN}$(mJ) | $E_{SNN}$(mJ) | E(mJ) |
|---|---|---|---|---|---|---|---|---|---|
| **ANN-ANN** | 178.11M | 56.4 | - | 112.7 | - | - | 259.3 | - | 259.3 |
| **ANN-SNN (T=1)** | 158.05M | 31.0 | 28.5 | 58.3 | 3.7 | 4.2 | 133.7 | 12.2 | 146.0 |
| **ANN-SNN (T=3)** | 158.05M | 34.7 | 85.5 | 58.5 | 10.9 | 11.4 | 133.8 | 35.3 | 169.1 |
| **ISTA Adapter** | 0.32M | 0.16 | - | 0.31 | - | - | 0.72 | - | 0.72 |
| **TDA** | 72 | 0.0001 | - | 0.0002 | - | - | 0.0006 | - | 0.0006 |

[13] Z. Wu, J. Zheng, X. Ren, F.-A. Vasluianu, C. Ma, D. P. Paudel, L. Van Gool, and R. Timofte, "Single-Model and Any-Modality for Video Object Tracking," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2024, pp. 19 156–19 166.

[14] X. Chen, J. Pan, and J. Dong, "Bidirectional Multi-Scale Implicit Neural Representations for Image Deraining," Apr 2024. [Online]. Available: http://arxiv.org/abs/2404.01547

[15] C. Cao, X. Fu, Y. Zhu, Z. Sun, and Z.-J. Zha, "Event-Driven Video Restoration With Spiking-Convolutional Architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2023.

[16] Z. Liu, J. Wu, G. Shi, W. Yang, W. Dong, and Q. Zhao, "Motion-Oriented Hybrid Spiking Neural Networks for Event-Based Motion Deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3742–3754, May 2024.

[17] S. Xiao, Y. Li, Y. Kim, D. Lee, and P. Panda, "ReSpike: Residual Frames-based Hybrid Spiking Neural Networks for Efficient Action Recognition," Sep 2024. [Online]. Available: http://arxiv.org/abs/2409.01564

[18] X. Wang, Z. Wu, Y. Rong, L. Zhu, B. Jiang, J. Tang, and Y. Tian, "SSTFormer: Bridging Spiking Neural Network and Memory Support Transformer for Frame-Event based Recognition," Aug 2023. [Online]. Available: http://arxiv.org/abs/2308.04359

[19] D. Li, J. Li, X. Liu, Z. Zhou, X. Fan, and Y. Tian, "HDI-Former: Hybrid Dynamic Interaction ANN-SNN Transformer for Object Detection Using Frames and Events," Nov 2024.

[20] H. Sun, R. Liu, W. Cai, J. Wang, Y. Wang, H. Tang, Y. Cui, D. Yao, and D. Guo, "Reliable object tracking by multimodal hybrid feature extraction and transformer-based fusion," *Neural Networks*, vol. 178, p. 106493, Oct 2024.

[21] Y. Ji, Q. Zhao, Y. Liang, and J. Wu, "SNNPTrack: Spiking Neural Network Based Prompt for High-Accuracy RGBE Tracking," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2025, pp. 1–5.

[22] A. Aydin, M. Gehrig, D. Gehrig, and D. Scaramuzza, "A Hybrid ANN-SNN Architecture for Low-Power and Low-Latency Visual Perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5701–5711.

[23] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[24] X. Deng and P. L. Dragotti, "Deep Convolutional Neural Network for Multi-Modal Image Restoration and Fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct 2021.

[25] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Designing CNNs for Multimodal Image Restoration and Fusion via Unfolding the Method of Multipliers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5830–5845, Sep 2022.

[26] S. Li, B. Zhang, L. Fei, S. Zhao, and Y. Zhou, "Learning Sparse and Discriminative Multimodal Feature Codes for Finger Recognition," *IEEE Trans. Multimed.*, vol. 25, pp. 805–815, 2023.

[27] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov 2004.

[28] K. Gregor and Y. LeCun, "Learning Fast Approximations of Sparse Coding," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 399–406.

[29] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021.

[30] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Springer International Publishing, 2016, pp. 850–865.

[31] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2018, pp. 8971–8980.

[32] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate Tracking by Overlap Maximization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2019, pp. 4655–4664.

[33] M. Danelljan, L. Van Gool, and R. Timofte, "Probabilistic Regression for Visual Tracking," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2020, pp. 7181–7190.

[34] C. Mayer, M. Danelljan, D. Pani Paudel, and L. Van Gool, "Learning Target Candidate Association to Keep Track of What Not to Track," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct 2021, pp. 13 424–13 434.

[35] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning Spatio-Temporal Transformer for Visual Tracking," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct 2021, pp. 10 428–10 437.

[36] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "SwinTrack: A Simple and Strong Baseline for Transformer Tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2021, pp. 8122–8131.

[37] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "AiATrack: Attention in Attention for Transformer Visual Tracking," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer Nature Switzerland, 2022, vol. 13682, pp. 146–164.

[38] Y. Cui, C. Jiang, L. Wang, and G. Wu, "MixFormer: End-to-End Tracking with Iterative Mixed Attention," Mar 2022. [Online]. Available: http://arxiv.org/abs/2203.11082

[39] Y. Han, X. Yu, H. Luan, and J. Suo, "Event-Assisted Object Tracking on High-Speed Drones in Harsh Illumination Environment," *Drones*, vol. 8, no. 1, p. 22, Jan 2024.

[40] X. Wang, J. Huang, S. Wang, C. Tang, B. Jiang, Y. Tian, J. Tang, and B. Luo, "Long-term Frame-Event Visual Tracking: Benchmark Dataset and Baseline," Apr 2024. [Online]. Available: http://arxiv.org/abs/2403.05839

[41] Y. Zhu, X. Wang, C. Li, B. Jiang, L. Zhu, Z. Huang, Y. Tian, and J. Tang, "CRSOT: Cross-Resolution Object Tracking using Unaligned Frame and Event Cameras," Jan 2024. [Online]. Available: http://arxiv.org/abs/2401.02826

[42] P. Shao, T. Xu, Z. Tang, L. Li, X.-J. Wu, and J. Kittler, "TENet: Targetness entanglement incorporating with multi-scale pooling and mutually-guided fusion for RGB-E object tracking," *Neural Networks*, vol. 183, p. 106948, Mar 2025.

[43] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual Prompt Multi-Modal Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9516–9526.

[44] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, and W. Zhang, "OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 079–19 091.

[45] B. Cao, J. Guo, P. Zhu, and Q. Hu, "Bi-directional Adapter for Multimodal Tracking," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 927–935, Mar 2024.

[46] R. C. O'Reilly, R. Bhattacharyya, M. D. Howard, and N. Ketz, "Com-

plementary Learning Systems," *Cognitive Science*, vol. 38, no. 6, pp. 1229–1248, Aug 2014.

[47] D. Kumaran, D. Hassabis, and J. L. McClelland, "What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated," *Trends in Cognitive Sciences*, vol. 20, no. 7, pp. 512–534, Jul 2016.

[48] R. Zhao, Z. Yang, H. Zheng, Y. Wu, F. Liu, Z. Wu, L. Li, F. Chen, S. Song, J. Zhu, W. Zhang, H. Huang, M. Xu, K. Sheng, Q. Yin, J. Pei, G. Li, Y. Zhang, M. Zhao, and L. Shi, "A framework for the general design and computation of hybrid neural networks," *Nat Commun*, vol. 13, no. 1, p. 3427, Jun 2022.

[49] F. Liu, H. Zheng, S. Ma, W. Zhang, X. Liu, Y. Chua, L. Shi, and R. Zhao, "Advancing brain-inspired computing with hybrid neural networks," *Natl. Sci. Rev.*, vol. 11, no. 5, p. nwae066, Apr 2024.

[50] Q. Shi, F. Liu, H. Li, G. Li, L. Shi, and R. Zhao, "Hybrid neural networks for continual learning inspired by corticohippocampal circuits," *Nat Commun*, vol. 16, no. 1, p. 1272, Feb 2025.

[51] Y. Wu, B. Shi, Z. Zheng, H. Zheng, F. Yu, X. Liu, G. Luo, and L. Deng, "Adaptive spatiotemporal neural networks through complementary hybridization," *Nat Commun*, vol. 15, no. 1, p. 7355, Aug 2024.

[52] J. Zhang, B. Dong, H. Zhang, J. Ding, F. Heide, B. Yin, and X. Yang, "Spiking Transformers for Event-based Single Object Tracking," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2022, pp. 8791–8800.

[53] J. Cao, X. Zheng, Y. Lyu, J. Wang, R. Xu, and L. Wang, "Chasing Day and Night: Towards Robust and Efficient All-Day Object Detection Guided by an Event Camera," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024, pp. 9026–9032.

[54] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov 2009.

[55] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan 2011.

[56] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and M. Prato, "Deep unfolding of a proximal interior point method for image restoration," *Inverse Problems*, vol. 36, no. 3, p. 034005, Feb 2020.

[57] C. Mou, Q. Wang, and J. Zhang, "Deep Generalized Unfolding Networks for Image Restoration," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2022, pp. 17 378–17 389.

[58] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event Enhanced High-Quality Image Recovery," in *European Conference on Computer Vision*. Springer, 2020.

[59] X. Deng and P. L. Dragotti, "Deep Coupled ISTA Network for Multi-Modal Image Super-Resolution," *IEEE Trans. on Image Process.*, vol. 29, no. 10, pp. 1683–1698, Oct 2020.

[60] L. Yu, B. Wang, X. Zhang, H. Zhang, W. Yang, J. Liu, and G.-S. Xia, "Learning to Super-Resolve Blurry Images With Events," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10 027–10 043, Aug 2023.

[61] S. Liu and P. L. Dragotti, "Sensing Diversity and Sparsity Models for Event Generation and Video Reconstruction from Events," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–16, 2023.

[62] X. Deng, E. Liu, S. Li, Y. Duan, and M. Xu, "Interpretable Multi-Modal Image Registration Network Based on Disentangled Convolutional Sparse Coding," *IEEE Trans. Image Process.*, vol. 32, pp. 1078–1091, 2023.

[63] J. Xu, X. Deng, and M. Xu, "Revisiting Convolutional Sparse Coding for Image Denoising: From a Multi-Scale Perspective," *IEEE Signal Process. Lett.*, vol. 29, pp. 1202–1206, 2022.

[64] C. Zhou, L. Yu, Z. Zhou, Z. Ma, H. Zhang, H. Zhou, and Y. Tian, "Spikingformer: Spike-driven Residual Learning for Transformer-based Spiking Neural Network," May 2023. [Online]. Available: http://arxiv.org/abs/2304.11954

[65] Q. Su, Y. Chou, Y. Hu, J. Li, S. Mei, Z. Zhang, and G. Li, "Deep Directly-Trained Spiking Neural Networks for Object Detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct 2023, pp. 6532–6542.

[66] Y. Hu, L. Deng, Y. Wu, M. Yao, and G. Li, "Advancing Spiking Neural Networks towards Deep Residual Learning," Mar 2023.

[67] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb 2014, pp. 10–14.