Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhancing Object Tracking via Spatio-Temporal Collaboration in Frame-Event Networks

Nan Wang
Beijing Institute of Technology
Yong Song
yongsong@bit.edu.cn
Beijing Institute of Technology
Li Wang
Beijing Institute of Control Engineering
Yingbo He
Beijing Institute of Control Engineering
Shaoxing Wu
Beijing Institute of Technology
Ya Zhou
Beijing Institute of Technology
Teng Luo
Beijing Institute of Technology
Shuang Wang
Beijing Institute of Technology
Linjun Zeng
Beijing Institute of Technology

Research Article

Keywords: Visual object tracking, Event cameras, Multimodal fusion, Spatio-temporal features , Adaptive spiking neurons

Posted Date: July 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9143315/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on October 23rd, 2026. See the published version at https://doi.org/10.1007/s00530-026-02749-1.

Enhancing Object Tracking via Spatio-Temporal Collaboration in Frame-Event Networks

Nan Wang[1], Yong Song[1*], Li Wang[2], Yingbo He[2], Shaoxing Wu[1], Ya Zhou[1], Teng Luo[1], Shuang Wang[1], Linjun Zeng[1]

[1*]School of Optics and Photonics, Beijing Institute of Technology, Beijing, 100081, China.
[2]Beijing Institute of Control Engineering, Beijing, 100190, China.

*Corresponding author(s). E-mail(s): yongsong@bit.edu.cn;
Contributing authors: 3120240652@bit.edu.cn; wupeng@163.com; ybhe86@126.com; wushaoxing@bit.edu.cn; zhouya@bit.edu.cn; 835676321@qq.com; 16601297106@163.com; 13654251879@163.com;

Abstract

Object tracking in challenging conditions, such as high-speed motion and extreme lighting, demands robust perception capabilities. Event cameras, with their high temporal resolution and dynamic range, complement conventional frame-based sensors. Leveraging both frame and event modalities, we propose the Spatio-Temporal Collaboration Network (STC-Net) to deeply integrate spatial and temporal information. Our approach introduces an Adaptive Spiking Neuron (ASN) module to process raw event streams, preserving high temporal resolution while filtering noise. A High-order Spatio-Temporal Fusion (HSTF) module aligns and fuses RGB semantics with event dynamics. Experiments on the FE108 and VisEvent datasets demonstrate state-of-the-art performance, outperforming existing RGB-Event trackers by 3.7% in RPR and 2.0% in RSR on the FE108 benchmark. Ablation studies validate the effectiveness of each key component. Project will be available at https://github.com/wangnan0623/STC-Net.

Keywords: Visual object tracking · Event cameras · Multimodal fusion · Spatio-temporal features · Adaptive spiking neurons

1 Introduction

Object tracking is a fundamental task in the field of computer vision [1–4], which aims to estimate the position and size of a given target in subsequent video frames based on its initial bounding box. In recent years, the advancement of deep learning methods has greatly propelled the development of visual tracking, revealing substantial potential and enabling a wide range of applications, such as intelligent video surveillance, robotics, and autonomous driving. However, most existing deep learning-based trackers [5–9] rely on conventional frame-based cameras as sensing devices and often struggle in challenging scenarios, including low illumination, fast motion, and background clutter.

To address these limitations, some researchers resort to bio-inspired event cameras for more accurate and robust object tracking. Different from conventional frame-based cameras which capture the absolute light intensity in a synchronous manner, event cameras measure pixel-level log-intensity light changes and output a stream of asynchronous events to represent visual information. This unique mechanism allows event cameras to capture motion information with microsecond-level latency, a high dynamic range (up to 140 dB), and low power consumption [10, 11]. Moreover, since event cameras record changes in light intensity rather than absolute intensity, they can work in extreme conditions such as low light and overexposure, where frame-based cameras fail to capture complete information.

Despite benefits, event cameras struggle to capture slow-moving or static objects and often lack fine-grained texture information, which is crucial for high-performance tracking. Therefore, integrating frame (RGB) and event cameras has become a natural direction toward more reliable object tracking. Early studies [12–14] validated the feasibility of improving tracking performance through RGB-Event (RGB-E) fusion. Subsequent research has largely focused on designing multimodal feature fusion modules to enhance interaction between the two modalities. For instance, Zhang et al. [15] introduced a cross-domain feature integrator to effectively and adaptively fuse visual cues from both the frame and event domains. Wang et al. [16] built a simple but effective cross-modality transformer module, achieving more effective and robust tracking. Zhang et al. [17] designed a event-guided alignment framework that performs cross-modality and cross-frame-rate alignment simultaneously.

However, most of these methods overlook the temporal information in event data and fail to leverage its high temporal resolution. Moreover, they typically establish only shallow feature-level fusion, lacking deeper, high-level synergistic relationships between the two modalities.

In this work, we propose a Spatio-Temporal Collaboration Network (STC-Net) that fully exploits both spatial and temporal cues from RGB-Event data for robust tracking in challenging scenarios. Our main contributions are summarized as follows:

- We propose STC-Net, a tracking framework that deeply integrates spatial appearance from RGB and high-temporal-resolution dynamics from events through a carefully designed two-pronged approach.
- We propose an Adaptive Spiking Neuron (ASN) Module for event stream processing, which dynamically adjusts firing thresholds and membrane time constants to filter noise and enhance motion-related events, effectively preserving high-resolution temporal dynamics.
- We design a High-order Spatio-Temporal Feature Fusion (HSTF) Module that facilitates deep cross-modal interaction in frequency, spatial, and channel dimensions, enabling complementary fusion of rich appearance semantics from RGB and precise temporal dynamics from events.
- We conduct extensive experiments on two public RGB-Event tracking benchmarks, FE108 and VisEvent, demonstrating that our method outperforms state-of-the-art trackers. Ablation studies further validate the effectiveness of each proposed component.

2 Related work

2.1 Visual Object Tracking

Most of the current trackers are developed based on RGB cameras and track the target object frame by frame. These methods can be divided into two categories: correlation filter (CF)-based trackers [18–20] and deep learning-based trackers [5–8, 21, 22]. CF-based approaches learn a discriminative filter from the initial target appearance and apply it to subsequent frames through correlation operations. In contrast, deep trackers typically employ Siamese or Transformer architectures to extract features from template and search regions, and estimate similarity maps via cross-correlation or attention mechanisms [2, 23, 24].

Even though traditional CNN-based RGB trackers or novel Transformer-based trackers have already achieved impressive accuracy in well-conditioned environments, their performance degrades significantly under challenging conditions such as low illumination, high-speed motion,

and complex background clutter. This performance degradation stems largely from the fundamental limitations of conventional frame-based sensing: limited dynamic range, susceptibility to motion blur, and fixed, low temporal sampling rates.

To overcome these issues, researchers have begun to explore event-based tracking, leveraging the unique characteristics of bio-inspired event cameras [10, 25, 26]. Unlike frame-based sensors, event cameras capture brightness changes asynchronously at microsecond temporal resolution, offering high dynamic range (140 dB) and immunity to motion blur. Early event-based trackers, such as the mean-shift approach by Litzenberger et al. [27], applied clustering techniques directly to the event stream for real-time vehicle and pedestrian tracking. Others have drawn inspiration from spiking neural networks (SNNs) [4]. For example, Pikatkowska et al.[28] introduced Gaussian mixture models into the clustering algorithm for multiple persons tracking in the occurrence of high occlusions. Chen et al. [29] proposed a synchronous representation strategy to encode spatio-temporal information of asynchronous events. Yao et al. [30] introduced a temporal attention mechanism in SNN to weight event frames according to their importance, addressing the sparsity and uneven distribution of event data.

Despite these advances, purely event-based methods often struggle when event data are too sparse (e.g., slow-moving or static targets) or lack rich appearance semantics, leading to unreliable localization. This has motivated the growing interest in RGB-Event (RGB-E) multi-modal tracking, which seeks to combine the complementary strengths of both sensing modalities. However, considering the vast distribution gap between RGB and Event data, designing an effective cross-modal information fusion strategy remains a challenge for RGB-E tracking.

2.2 Spiking Neuron Model for Event Processing

Spiking Neural Networks (SNNs) have emerged as a powerful paradigm for processing asynchronous event streams due to their biological plausibility and inherent temporal dynamics. Unlike traditional Artificial Neural Networks (ANNs) that operate on static, frame-based representations, SNNs process information as a sequence of spikes over time, making them a natural fit for event data processing.

SNNs are built upon spiking neuron models, with the Leaky Integrate-and-Fire (LIF) [31] being one of the most widely adopted model due to its balance between biological plausibility and computational efficiency. A key property of LIF neurons is their temporal memory: they integrate incoming spikes into a membrane potential that decays over time, and only emit an output spike when the accumulated potential exceeds a threshold. This inherent property is highly advantageous for processing event streams. Events originating from a moving target tend to form dense spatio-temporal clusters, leading to rapid potential accumulation in a LIF neuron and subsequent firing. In contrast, sparse noise events or those from static backgrounds are less likely to trigger a response. Consequently, a single layer of LIF neurons can act as an effective filter, enhancing signal and suppressing noise based on temporal frequency.

However, standard LIF neurons employ fixed parameters—the firing threshold and membrane time constant—which struggle to accommodate the widely varying event statistics (e.g., density, motion speed) across different scenes. An improperly initialized threshold may miss true signals or amplify noise, while a fixed time constant cannot optimally balance integration for both fast and slow motion.

Recent works have explored adaptive mechanisms to enhance the flexibility of spiking neurons. Wu et al. [32] proposed BD-LIF neurons, which are capable of emitting both +1 and -1 spikes while dynamically adjusting the firing threshold based on membrane potential depolarization rate, achieving competitive accuracy in object detection. Fang et al. [33] proposed the Parametric LIF (PLIF) neuron, making τ learnable during training to reduce sensitivity to initial values. Zhang et al. [34] introduced STNet, which dynamically adjusts the firing threshold based on global spatial feature statistics. Nagaraj et al. [35] demonstrated that a lightweight single-layer spiking architecture with fixed parameters can separate objects by motion speed without extensive training. Most recently, Zhang et al. [36] proposed BA-LIF, which generates a spatial attention

map from inputs to dynamically adjust τ—assigning small values (slow decay) to salient regions to preserve information, and large values (fast decay) to background to accelerate noise leakage.

Despite these advances, existing adaptive mechanisms still exhibit certain limitations. PLIF [33] introduces layer-wise learnable constants but lacks input-dependent adaptation, making it unable to respond to per-frame variations. STNet [34] achieves dynamic threshold adjustment but relies on global spatial statistics, potentially overlooking fine-grained local patterns critical for distinguishing target from background. BA-LIF [36] incorporates spatial attention for adaptive decay, yet primarily focuses on spatial modulation without jointly considering temporal characteristics such as event density and motion variance. These observations motivate our design of the ASN module, which comprehensively integrates both threshold adaptation based on local event statistics and decay adaptation via channel attention, enabling more fine-grained and context-aware filtering of event data.

2.3 Multi-modal Feature Fusion Strategies

Effective integration of RGB and Event data is crucial for achieving robust tracking in challenging conditions. While theoretical frameworks [37] explain the superiority of multi-modal sensing, designing a practical fusion strategy remains non-trivial due to two core challenges: (1) the inherent heterogeneity between dense RGB appearance semantics and sparse, high-temporal-resolution event streams, and (2) the difficulty of modeling deep, synergistic relationships beyond simple feature combination. Existing approaches can be categorized by the stage at which modalities are integrated:

Early Fusion (Pixel-Level). This strategy concatenates raw or minimally processed inputs (e.g., RGB frames with event accumulations) before feeding them into a single network [38–41]. While computationally efficient, it requires precise spatial alignment. More critically, it forces the network to learn directly from the low-level disparity between intensity values and event counts, potentially obscuring the distinct characteristics of each modality. Furthermore, by collapsing asynchronous events into synchronous image channels, it discards the fine-grained temporal structure that is a key advantage of event-based sensing.

Late Fusion (Decision-Level). In this paradigm, each modality is processed through independent pipelines; their final outputs (e.g., confidence maps) are combined via weighted averaging or learnable aggregation [42–44]. Although modular, this approach only interacts at the decision level, failing to exploit the rich, complementary information available at intermediate feature levels where RGB semantics and event-based motion cues could be most mutually reinforcing.

Middle Fusion (Feature-Level). As the dominant paradigm in recent works [15–17, 45–48], middle fusion integrates intermediate features from separate or shared backbones. This balances modality-specific processing with cross-modal interaction. The design of the fusion operation is pivotal. Common choices include concatenation, summation, and attention mechanisms [49]. Systematic studies in related domains [50] and empirical evaluations in RGB-Event tracking [16] suggest that feature-level fusion, particularly via concatenation, offers a good trade-off, as it preserves complete modality-specific information.

However, prevalent middle-fusion implementations suffer from two significant limitations. First, they predominantly rely on first-order or simplistic second-order interactions (e.g., concatenation, basic attention). These mechanisms treat the two feature sets as largely independent, lacking the expressiveness to capture complex, high-order synergies across spatial, temporal, and channel dimensions. Second, they almost universally rely on event data converted into synchronous pseudo-frames for compatibility with standard backbones, thereby sacrificing the native microsecond temporal resolution of the raw event stream.

Our work directly confronts these shortcomings. We introduce an adaptive spiking neuron module to process event streams, preserving their high-temporal-resolution dynamics, and propose a high-order spatio-temporal feature fusion module to enable deep, structured cross-modal interaction. This two-pronged approach exploits the complementary strengths inherent in RGB-Event data for visual tracking.

3 Proposed method

3.1 Network Overview

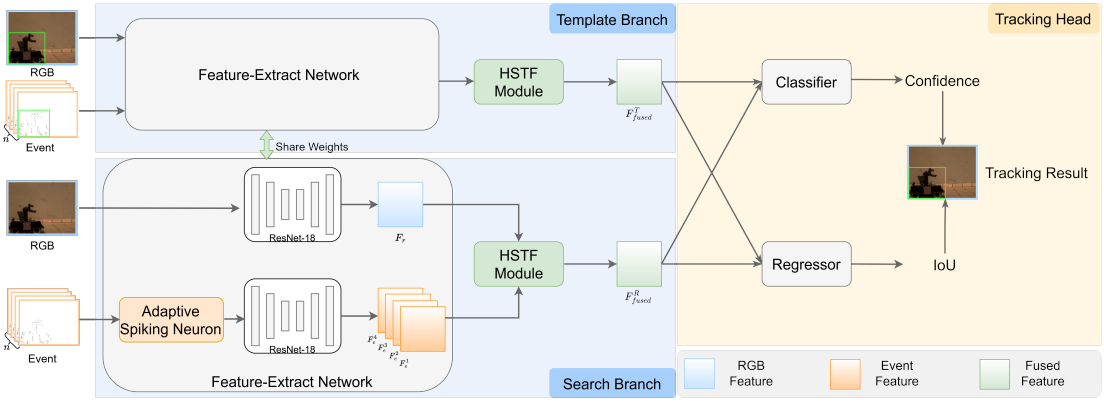


Fig. 1: Overall architecture of the proposed Spatio-Temporal Collaboration Network (STC-Net). The event stream is first processed by an adaptive spiking neuron module to suppress noise and enhance motion-related event clusters. Both the RGB image and the enhanced event representation are then fed into a shared ResNet-18 backbone for feature extraction. The resulting dual-modal features are fused through the proposed High-order Spatio-Temporal Feature (HSTF) module, enabling deep cross-modal interaction. The fused features are subsequently passed to a classifier and a regressor to predict target confidence scores and bounding-box offsets, respectively. The best-scored proposal is selected as the tracking output for the current frame, and the process is repeated sequentially throughout the video.

As shown in Figure 1, the overall architecture of our proposed Spatio-Temporal Collaboration Network (STC-Net) consists of three main components: the feature extractor (including backbone, adaptive spiking neuron, and HSTF module), the target classifier, and the bounding-box regressor. Following the DiMP [51] framework, the template and search branches share the same architecture and weights. Each branch takes an RGB image and its corresponding aggregated events as input. The event input first passes through an adaptive spiking neuron module, which suppresses noise in the event stream while enhancing event clusters triggered by moving objects. Subsequently, both the RGB and processed event data are fed into a shared backbone (ResNet-18) for feature extraction. The resulting dual-modal features are then passed to the HSTF module to achieve deep cross-modal interaction and fusion. Both the target classifier and the bounding-box regressor receive the fused features from the feature extractor. Given a template set of fused features and their corresponding target boxes, the model predictor learns the weights of the target classifier. These weights are then applied to the features extracted from the search branch to predict target confidence scores. Meanwhile, the bounding-box regressor estimates the IoU between the ground-truth and predicted bounding boxes.

The adaptive spiking neuron (ASN) module and the high-order spatio-temporal feature fusion (HSTF) module serve as the two core innovations of our approach, aiming to address the limitations of temporal dynamics and shallow fusion. The ASN module processes event input and leverages the memory mechanism of spiking neurons to adaptively retain dense event clusters triggered by moving objects while filtering out irrelevant noise. The HSTF module collaboratively utilizes high-order information across frequency, spatial, and channel dimensions, enabling deep interaction and complementary fusion between RGB and event features.

3.2 Events Representation

Different from visible sensors that capture the global information by synchronously recording the absolute intensity of all pixels in a frame-wise manner, event cameras respond asynchronously to per-pixel log-intensity changes. Each pixel independently emits a discrete event when the logarithmic brightness (I) change exceeds a predefined threshold (C) [11]:

$$I(x, y, t) - I(x, y, t - \Delta t) \geq pC \tag{1}$$

where $p \in \{+1, -1\}$ denotes the polarity of the brightness change, and Δt is the tiny time period since the last event at location $(x, y)^T$.

The output of an event camera can be formally described as a spatio-temporal stream of raw events ε:

$$\varepsilon = \{e_i = (x_i, y_i, t_i, p_i) | i \in I\}, I = [1, 2, ..., n] \tag{2}$$

where (x_i, y_i) represent the spatial coordinate where the event emits, t_i is the microsecond-resolution timestamp, and p_i indicates the polarity. Each event thus carries 2D spatial, 1D temporal, and 1D polarity information (often summarized as 2S–1T–1P).

To interface with synchronous deep learning frameworks, raw events are commonly aggregated over a fixed time window T into image-like representations. The resulting event frame, capturing events within a window $[t, t + T)$, is defined as:

$$F_T(t) = \mathcal{Q}\left(e_i \mid t_i \in [t, t + T)\right) \tag{3}$$

where $\mathcal{Q}(\cdot)$ denotes an aggregation function—such as non-polarity accumulation, polarity-aware summation, or logic-based merging—that maps the asynchronous event stream to a synchronous 2D grid.

In this work, to preserve the high temporal resolution of events while maintaining alignment with RGB frames, we partition the exposure period of each RGB image into n sub-intervals. Within each sub-interval, events are accumulated into an event frame, with polarity information retained to encode the direction of brightness changes. This representation provides a balance between temporal granularity and computational efficiency. Crucially, by preserving multiple frames within an RGB exposure, it retains more temporal information than a single pseudo-frame, providing richer sequential input for our subsequent Adaptive Spiking Neuron (ASN) module and enabling more effective multi-modal fusion.

3.3 Adaptive Spiking Neuron Module

Directly utilizing raw asynchronous and sparse event streams for object tracking presents two main difficulties. First, event cameras are susceptible to noise. The resulting noise events are indistinguishable from true signal events by magnitude, making filtering difficult. Second, while events inherently encode high-temporal-resolution motion information, effectively extracting and utilizing these temporal cues—especially to separate moving objects from static background or noise—remains challenging.

To address this, we propose an Adaptive Spiking Neuron (ASN) module that dynamically adjusts both firing threshold and membrane decay based on input characteristics. Instead of using a full SNN architecture, we only use a single layer of ASN to process raw event data. The membrane dynamics of standard LIF model can be defined as [33]:

$$\begin{cases} \text{Charge:} & V(t) = \left(1 - \dfrac{1}{\tau}\right) V(t-1) + \dfrac{1}{\tau} X(t) \\ \text{Fire:} & S(t) = \Theta(V(t) - V_{th}) \\ \text{Reset:} & H(t) = (1 - S(t))V(t) + S(t)V_{reset} \end{cases} \tag{4}$$

where $V(t)$ and $V(t-1)$ represent the membrane potential of the neuron at time t and $t-1$, $X(t)$ represents the input to neuron at time t, $H(t)$ represents the membrane potential after the trigger of a spike, and $\Theta(\cdot)$ denotes the Heaviside step function. The accumulated membrane potential

decays over time and the decay rate is controlled by the decay factor τ. The integration progress $\frac{1}{\tau}X(t)$ makes the neuron able to remember current input information, while the leakage progress $(1-\frac{1}{\tau})V(t-1)$ can be seen as forgetting some information from the past. If the accumulated membrane potential $V(t)$ exceeds the threshold V_{th} at any point, the neuron emits an output spike and resets its membrane potential (here we assume the reset membrane potential $V_{reset}=0$).

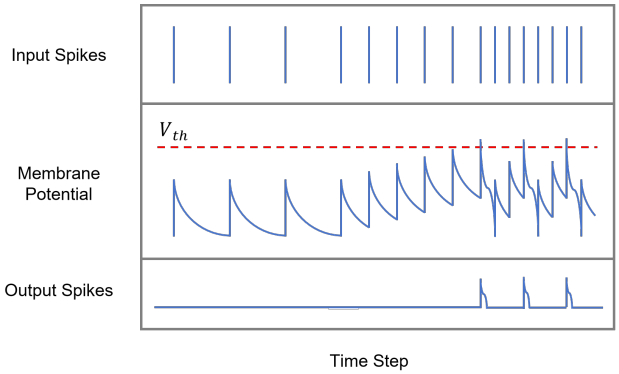


Fig. 2: Membrane potential dynamics of a spiking neuron. The red dashed line indicates the firing threshold of the neuron. As the input spike frequency increases over time steps (illustrated by the denser spike train), the neuron's membrane potential accumulates more rapidly. Under a fixed threshold, a higher input frequency leads to a greater likelihood of reaching the threshold and emitting an output spike, which aligns with our description of how spiking neurons can effectively respond to motion-related event clusters.

As shown in Figure 2, the behavior of a spiking neuron is governed by its membrane dynamics: when stimulated by a single input spike, its membrane potential rises sharply and then gradually decays toward the resting state. However, when the neuron receives a stream of spikes in close succession, the membrane potential accumulates continuously until reaching a firing threshold, at which point an output spike is emitted. In other words, a neuron only generates an output spike when the input event frequency exceeds a certain threshold. In raw event streams, events caused by a moving target tend to cluster within a small spatial region and exhibit higher temporal frequency, whereas events from static background variations (e.g., due to camera shake or illumination changes) appear with lower frequency, and noise events are entirely irregular. By exploiting this property of spiking neurons in conjunction with the frequency characteristics of event data, our ASN module can effectively distinguish target-related events from background and noise based on their temporal firing patterns, thereby obtaining a cleaner and more informative event representation. As shown in Figure 3, the adjust process of V_{th} and τ can be defined as:

The adaptive firing threshold V_{th} at time step t is computed by combining three event characteristics:

$$V_{th} = V_{th}^{\text{base}} + \sigma\left(\omega_d \cdot d + \omega_m \cdot \sigma + \omega_c \cdot \Delta\right) \tag{5}$$

where V_{th}^{base} is the base threshold, d is the event density, σ is the local motion pattern variance, Δ is the temporal change between consecutive frames, and ω_d, ω_m, ω_c are learnable weights. $\sigma(\cdot)$ denotes the sigmoid activation function.

Instead of using a fixed τ, we compute an adaptive decay factor β for each channel. This formulation arises from the relationship between the decay factor β and the membrane time constant τ in the LIF model: $\beta = 1 - \frac{1}{\tau}$. Directly optimizing τ would pose numerical stability issues, as τ appears in the denominator and small values could lead to gradient explosion. By optimizing β instead, we maintain stability while preserving the dynamic decay behavior. Specifically, $\beta \in (0,1)$ corresponds to $\tau \in (1,\infty)$, ensuring stable membrane potential updates. The adaptive β is computed as:

$$\beta = \sigma\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot [\mathcal{A}(E(t)), \mathcal{M}(E(t))]\right)\right) \tag{6}$$

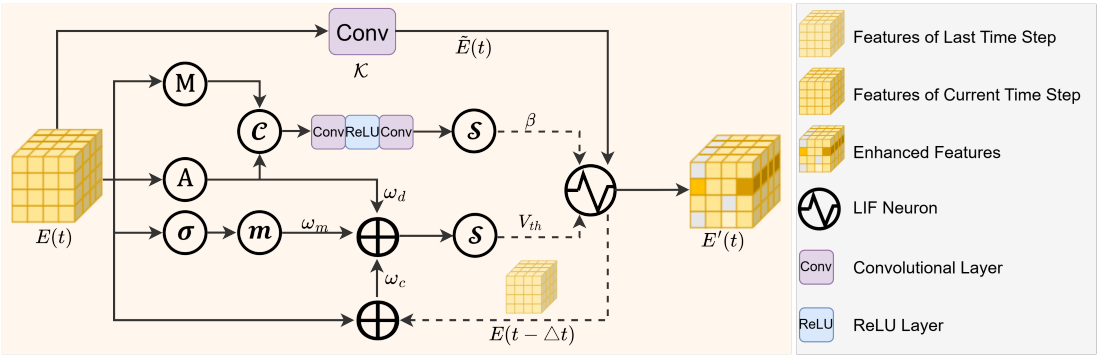


Fig. 3: Architecture of the proposed Adaptive Spiking Neuron (ASN) module. The input event stream $E(t)$ is first processed by a 3×3 spatial weighting kernel $\mathcal{K}$ (weights: center 0.2, periphery 0.1) to aggregate local neighborhood information. The input $E(t)$ feeds into two adaptive branches: (1) the threshold adaptation branch computes a dynamic firing threshold V_{th} based on event density d, motion variance σ, and temporal change Δ; (2) the decay adaptation branch generates an adaptive decay factor β via a channel attention mechanism that processes concatenated average- and max-pooled features. The adaptive LIF neuron integrates these components to generate a gating signal $S(t)$, which is then element-wise multiplied with the spatially weighted input $\tilde{E}(t)$ to produce the final output $E'(t) = \tilde{E}(t) \odot S(t)$. This operation retains motion-related features while suppressing noise.

where $\sigma(\cdot)$ is the sigmoid function, W_1 and W_2 are learnable linear transformations, $\mathcal{A}$ and $\mathcal{M}$ denote the average pooling and max pooling, and $[\cdot,\cdot]$ denotes concatenation. The adaptive membrane dynamics then become:

$$V^{(t)} = \beta^{(t)} \cdot V^{(t-1)} + (1 - \beta^{(t)}) \cdot \tilde{E}(t) \tag{7}$$

where $\tilde{E}(t) = \mathcal{K} * E(t)$ is the input after 3×3 spatial neighborhood weighting with kernel $\mathcal{K}$ initialized to emphasize central events. Specifically, we employ a 3×3 convolutional kernel with central weight 0.2 and peripheral weights 0.1 for each of the eight neighboring pixels. This design is motivated by the observation that moving objects trigger events not at a single fixed pixel, but across a small spatial region over consecutive time steps. By aggregating inputs from the local neighborhood, our architecture effectively captures temporal correlations between adjacent events, enabling robust detection of motion patterns that would otherwise be missed by per-pixel processing. This approach enhances the network's ability to isolate events belonging to fast-moving objects while maintaining spatial specificity.

Our ASN module employs only a single layer of spiking neurons to capture the temporal information present in event data. This mitigates the energy and latency overhead of the neural architecture while maintaining the ability to filter noise and enhance motion-related events through adaptive threshold and membrane dynamics.

3.4 High-order Spatio-Temporal Fusion Module

Traditional fusion methods often employ simple operations such as concatenation, addition, or second-order attention mechanisms, which fail to fully exploit the synergistic relationships between RGB and event modalities. To overcome this limitation, we propose a High-order Spatio-Temporal Feature (HSTF) fusion module that systematically explores high-order interactions across both spatial and channel dimensions, enabling deep complementary integration of appearance semantics from RGB and temporal dynamics from events.

The ASN module processes event streams in their native temporal format and outputs 5D features of shape $[T, B, C, H, W]$. However, directly propagating this temporal dimension through the entire network would incur significant computational cost with limited performance gains.

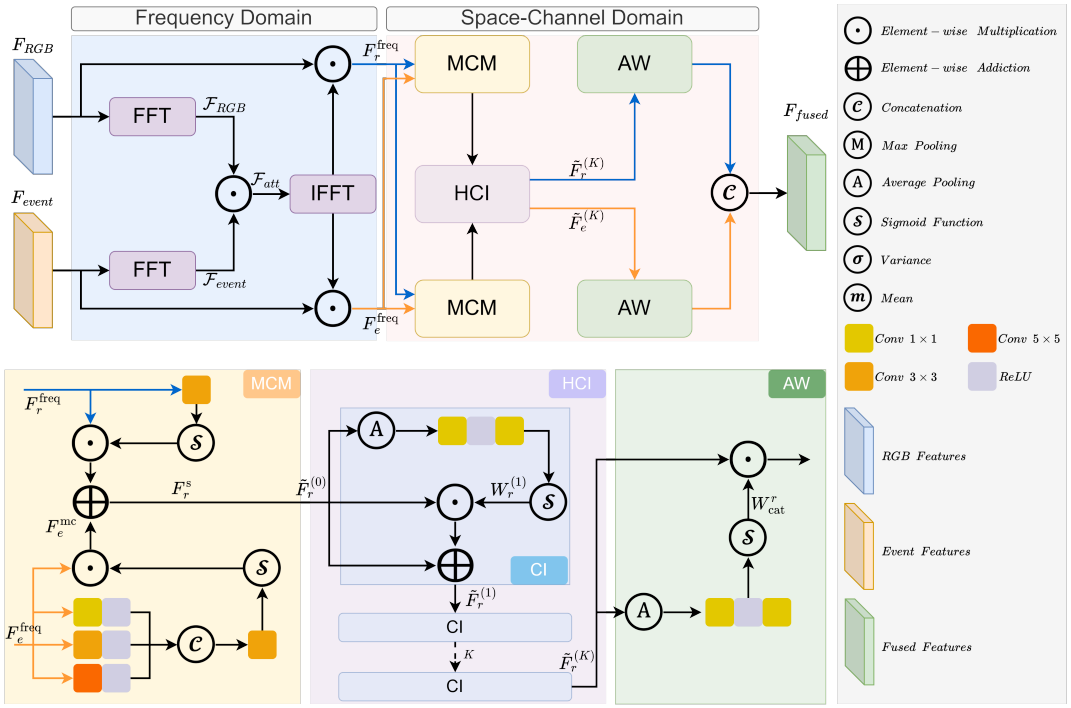


Fig. 4: Detailed architecture of our proposed High-order Spatio-Temporal Feature Fusion (HSTF) Module. RGB and event features are first processed by frequency-domain interaction via FFT to capture global correlations. The frequency-enhanced features then pass through a Multi-Context Modulation (MCM) block with multi-scale kernels for spatial interaction, followed by a High-order Channel Interaction (HCI) block with recursive squeeze-and-excitation to model channel dependencies. Finally, an Adaptive Weight (AW) block learns channel-wise fusion weights before concatenation.

To enable efficient yet effective fusion with RGB features, we collapse the temporal dimension via averaging before feeding event features into the HSTF module, yielding a 4D tensor of shape $[B, C, H, W]$. This design choice preserves the motion-enhanced representations learned by ASN while ensuring compatibility with the subsequent fusion operations.

As illustrated in Figure 4, the HSTF module operates on multi-scale features extracted from the backbone network. For each scale level (the output of Layer2 and Layer3 of the ResNet-18), the module processes RGB features $F_{RGB} \in \mathbb{R}^{B\times C\times H\times W}$ and event features $F_{event} \in \mathbb{R}^{B\times C\times H\times W}$ through a two-stage pipeline:

1. **First-order Frequency-domain Interaction**: Establishing initial cross-modal correlations through frequency-domain multiplication.
2. **High-order Spatial-Channel domain Interaction**: Enhancing spatial synergistic relationships via multi-scale contextual modulation and modeling inter-dependencies across channels.

Frequency-domain Interaction: Several works have explored the frequency domain in computer vision [3, 47, 52]. Inspired by the frequency-domain equivalence of spatial correlations, we transform RGB and event features into the frequency domain using 2D Fast Fourier Transform (FFT) to establish efficient first-order interactions. Given RGB features F_{RGB} and event features F_{event}, their frequency representations are obtained as:

$$\mathcal{F}_{RGB} = \mathrm{FFT}(F_{RGB}), \quad \mathcal{F}_{event} = \mathrm{FFT}(F_{event}) \tag{8}$$

The cross-modal correlation in frequency domain is computed through element-wise multiplication:

$$\mathcal{F}_{\text{att}} = \mathcal{F}_{RGB} \odot \mathcal{F}_{event} \tag{9}$$

where $\odot$ denotes Hadamard product.

The resulting attention map is transformed back to spatial domain via inverse FFT (IFFT) and normalized. The frequency-enhanced features are then obtained by element-wise multiplying the original features with this attention map. This frequency-domain interaction efficiently captures global spatial correlations while avoiding the computational overhead of traditional matrix multiplication operations.

High-order Spatial-Channel domain Interaction: Building upon the frequency-enhanced features, we design a Multi-Context Modulation (MCM) block for spatial interactions and a High-order Channel Interaction (HCI) block to capture complex feature dependencies. After this two stages, the RGB and event features are adaptively weighted and concatenated via an Adaptive Weight (AW) block to produce the final fused representation.

The MCM block employs a bidirectional guidance mechanism where each modality provides contextual information to modulate the other. For RGB feature modulation guided by event features:

$$F_e^{\text{mc}} = F_e^{\text{freq}} \otimes \sigma(\mathcal{M}(F_e^{\text{freq}})) \tag{10}$$

$$F_r^{\text{s}} = F_r^{\text{freq}} \otimes \sigma(F_r) + F_e^{\text{mc}} \tag{11}$$

where $\mathcal{M}$ is a multi-context module that extracts features at three scales (kernel sizes 1, 3, and 5), and σ denotes the sigmoid function. This bidirectional modulation enables each modality to selectively emphasize regions that are informative in the complementary modality, fostering synergistic spatial relationships.

The HCI block extends the squeeze-and-excitation mechanism to high-order formulations to capture complex inter-dependencies across feature channels. For K-order channel interaction, we recursively apply squeeze-and-excitation attention with residual connections. For $k = 1, 2, \ldots, K$, the process for RGB features is defined as:

$$\tilde{F}_r^{(0)} = F_r^{\text{s}} \tag{12}$$

$$W_r^{(k)} = \sigma\left(\mathcal{F}_2\left(\text{ReLU}\left(\mathcal{F}_1\left(\text{GAP}(\tilde{F}_r^{(k-1)})\right)\right)\right)\right) \tag{13}$$

$$\tilde{F}_r^{(k)} = \tilde{F}_r^{(k-1)} \otimes W_r^{(k)} + F_r^{\text{s}} \tag{14}$$

where GAP is global average pooling, $\mathcal{F}_1$ and $\mathcal{F}_2$ are linear transformations, σ is the sigmoid activation, and $\otimes$ denotes element-wise multiplication. Event features undergo the same process.

To adaptively balance the contributions of RGB and event modalities, we apply an Adaptive Weight (AW) block to learn channel-wise fusion weights:

$$W_{\text{cat}}^r = \sigma\left(\text{FC}\left(\text{GAP}(\tilde{F}_r^{(k)})\right)\right) \tag{15}$$

$$W_{\text{cat}}^e = \sigma\left(\text{FC}\left(\text{GAP}(\tilde{F}_e^{(k)})\right)\right) \tag{16}$$

The weighted features are then concatenated to produce the final fused representation:

$$F_{\text{fused}} = \text{Concat}\left[W_{\text{cat}}^r \otimes \tilde{F}_r^{(k)},\ W_{\text{cat}}^e \otimes \tilde{F}_e^{(k)}\right] \tag{17}$$

This hierarchical fusion strategy progressively integrates complementary information from RGB and event modalities: frequency-domain interactions capture global correlations, spatial interactions model contextual relationships, and channel interactions encode complex inter-dependencies. The resulting fused features preserve both rich appearance semantics and precise temporal dynamics, enabling robust tracking in challenging scenarios.

4 Experiments

4.1 Datasets and Evaluation Metric

We evaluate our STC-Net on two realistic event-frame-based dataset: FE108 [15] and VisEvent [16]. The FE108 dataset contains 108 video sequences, which are divided into 76 training sequences and 32 testing sequences. It provides a variety types of challenging scenarios, such as low-illumination, overexposure, multiple objects and motion blur. Compared with FE108, VisEvent dataset contains 820 video sequences with more outdoor scenes. It also includes more complex scenarios, such as abrupt motion of the camera and cases where the target completely leaves the video sequence. Following [15, 17], we remove sequences that miss event data or have misaligned timestamps, leaving 268 sequences for training and 176 for testing.

To evaluate the tracking performance of our proposed network, two widely used metrics, success rate (SR) and precision rate (PR), are adopted to quantitatively compare the performance of different trackers. Both SR and PR are calculated based on the ratio of successfully tracked frames to the total number of frames. Specifically, SR is determined by whether the Intersection over Union (IoU) between the predicted bounding box and the ground-truth bounding box exceeds a certain threshold. In contrast, PR is assessed by whether the distance between the center points of the predicted and ground-truth bounding boxes is less than a given threshold. For SR, we calculate the area under curve (AUC) of an SR plot as representative SR (RSR). For PR, we used the PR score associated with a 20-pixel threshold as representative PR (RPR) [15].

4.2 Experimental Settings

We adopt the pretrained ResNet-18 as the backbone to extract frame and event features. Following [17, 51], the loss function is defied as:

$$L = \beta L_{cls} + L_{bb} \tag{18}$$

where L_{cls} is the target classification loss and L_{bb} is the bounding box regressor loss. β is set to 100.

Based on Python 3.8 and PyTorch 2.0 deep learning API, we trained our network for 60 epochs with a batch size of 24 using Adam optimizer with the default parameters. The initial learning rate for the feature extraction network, the ASN module, the HSTF module, the classifier, and the regressor are set to $1e-4$, $1e-5$, $1e-4$, $1e-3$, $1e-3$, respectively. The learning rate is adjusted by the CosineAnnealingLR strategy. Our network is run on a single Nvidia RTX 3090 GPU with 24G memory.

4.3 Comparison with State-of-the art Trackers

To comprehensively evaluate the effectiveness of our proposed STC-Net, we conduct extensive comparisons with seven state-of-the-art tracking methods on both FE108 and VisEvent datasets. The compared methods include conventional frame-based trackers which are extended to multi-modality and trackers designed for RGB-E dual-modal. These baselines represent the current state-of-the-art in both conventional frame-based tracking and emerging RGB-Event multimodal tracking, providing a comprehensive benchmark for evaluation.

FE108 Dataset: As shown in Table 1, our STC-Net achieves the best performance on the FE108 dataset with **89.0% RPR** and **57.8% RSR**, significantly outperforming all compared methods. Our method shows substantial improvements over traditional frame-based approaches. Specifically, STC-Net outperforms DiMP18 by **3.9%** in RPR and **0.7%** in RSR, demonstrating the advantage of incorporating event data. Compared to ATOM, we achieve a gain of **7.2%** in RPR and **2.3%** in RSR. The improvements are even more pronounced when compared to Siamese-based trackers: **55.5%** higher RPR than SiamRPN++ and **49.9%** higher RPR than SiamFC++. These results validate that event data provides critical complementary information, especially in challenging scenarios where frame-based cameras fail. STC-Net also outperforms the state-of-the-art RGB-Event tracker AFNet by **3.7%** in RPR and **2.0%** in RSR. This performance gain can be

attributed to our adaptive spiking neuron module that effectively preserves temporal dynamics and the high-order spatio-temporal fusion module that enables deeper cross-modal interactions. While AFNet focuses on aligning the two modalities and enhancing features via cross-correlation, our method goes a step further by explicitly modeling high-order interactions across frequency, spatial, and channel dimensions, enabling deeper synergistic fusion beyond alignment and basic feature enhancement.

Table 1: Experimental Results (**RPR** | **RSR**) on the FE108 Dataset. The Results of Baseline Trackers are Borrowed from FE108 [15] Benchmark

STC-Net(Ours)	**CLNet**[53]	**PrDiMP18**[54]	**SiamRPN++**[55]
89.0 | 57.8	55.5 | 34.4	80.5 | 53.0	33.5 | 21.8
ATOM[56]	**AFNet**[17]	**SiamFC++**[57]	**DiMP18**[51]
81.8 | 55.5	85.3 | 55.8	39.1 | 23.8	85.1 | 57.1

VisEvent Dataset: Figure 5 presents the precision and success plots on the larger and more challenging VisEvent dataset. Our STC-Net achieves **64.0%** overall RPR and **46.8%** RSR, surpassing the runner-up method by **3.1%** and **1.0%**, respectively. Our method's consistent superiority on this dataset demonstrates its generalization capability and robustness across diverse environments. The precision plot shows that STC-Net maintains high tracking accuracy even at strict threshold settings, indicating precise localization capability. The success plot reveals that our method achieves higher overlap ratios with ground truth bounding boxes, suggesting better target representation and scale estimation.

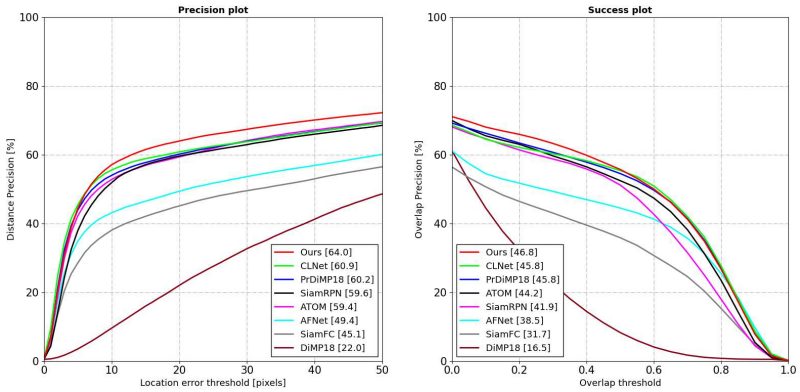


Fig. 5: Precision (left) and Success (right) plots on VisEvent dataset. Our method achieves the best performance on both metrics, outperforming state-of-the-art trackers.

4.4 Ablation Study

To validate the effectiveness of each key component in our proposed STC-Net, we conduct comprehensive ablation studies on the FE108 dataset. The ablation experiments are designed to isolate and quantify the contributions of the input modal, the Adaptive Spiking Neuron (ASN) module and the High-order Spatio-Temporal Feature (HSTF) fusion module. The experimental results are presented in Table 2.

Input Modal: The analysis of input modalities reveals several important insights. **Variant A** (RGB-only) achieves 74.7% RPR and 49.3% RSR, representing the baseline performance of conventional frame-based tracking. **Variant B** (Event-only) demonstrates the unique capability

Table 2: Ablation study on the FE108 dataset. Variants A-C compare different input combinations, while variants D-E progressively add our proposed ASN and HSTF modules to demonstrate their individual and combined effectiveness.

Variant	Input		Module		Results (%)	
	Frame	**Event**	**ASN**	**HSTF**	**RPR**	**RSR**
A	✓				74.7	49.3
B		✓			83.0	53.1
C	✓	✓			81.7	54.5
D	✓	✓	✓		86.0	56.7
E (Ours)	✓	✓	✓	✓	**89.0**	**57.8**

of event data, achieving 83.0% RPR and 53.1% RSR. Notably, event data provides superior precision (83.0% vs 74.7% RPR), outperforming RGB-only tracking by a significant margin. This advantage can be attributed to the characteristics of the FE108 dataset, which contains numerous sequences with low illumination and overexposure—scenarios where frame-based cameras struggle but event cameras excel due to their high dynamic range and immunity to lighting variations. These results highlight the complementary value of event data in challenging visual conditions. **Variant C** (simple concatenation of both modalities) achieves 81.7% RPR and 54.5% RSR. While its precision (81.7%) is slightly lower than Event-only (83.0%), it still substantially outperforms RGB-only tracking (74.7%), confirming the value of incorporating event data. More importantly, its success rate (54.5% RSR) surpasses both unimodal approaches (49.3% and 53.1%), demonstrating that even simple fusion can leverage complementary strengths from both modalities to achieve more robust tracking.

ASN Module: The effectiveness of our Adaptive Spiking Neuron module is clearly demonstrated by comparing **Variant D** (with ASN) against **Variant C** (without ASN). Variant D achieves 86.0% RPR and 56.7% RSR, representing improvements of +4.3% RPR and +2.2% RSR over Variant C. These gains validate the importance of specialized event processing, which stems from three key designs: (1) The learnable threshold mechanism dynamically adapts to event density and motion patterns; (2) The adaptive membrane time constants maintain temporal coherence across frames; (3) The spatial neighborhood weighting captures local correlations among adjacent events, complementing temporal integration for robust detection of fast-moving objects.

The dynamic threshold adjustment mechanism enables ASN to automatically regulate its sensitivity based on input characteristics. When a moving target generates dense event clusters, ASN appropriately raises the firing threshold to prevent over-activation and maintain selectivity. Conversely, when event signals are sparse due to slow motion or low contrast, ASN lowers the threshold to ensure meaningful events are not missed.

Similarly, the adaptive membrane time constants allow ASN to optimally balance temporal integration and forgetting. This is particularly beneficial when event frequency fluctuates significantly—during rapid acceleration or deceleration of moving objects. Classical LIF neurons with fixed parameters often either retain too much historical information (causing temporal blur) or forget too quickly (losing motion continuity). Our ASN module dynamically adjusts its memory characteristics to match the temporal dynamics of the input, resulting in more stable and coherent event representations.

As illustrated in Figure 6, we present a comparative visualization to validate the effectiveness of ASN module: (a) raw event stream input, (b) events filtered by classical LIF neurons with fixed parameters, and (c) events processed by our proposed ASN module. The visual comparison reveals two key advantages. First, ASN demonstrates superior noise suppression capability, effectively filtering out scattered noise while preserving dense event clusters triggered by moving targets—crucial for maintaining clean representations in real-world scenarios. Second, compared to fixed-parameter LIF neurons, our ASN exhibits adaptive behavior that better handles varying event patterns. By providing cleaner and more informative event representations, the ASN

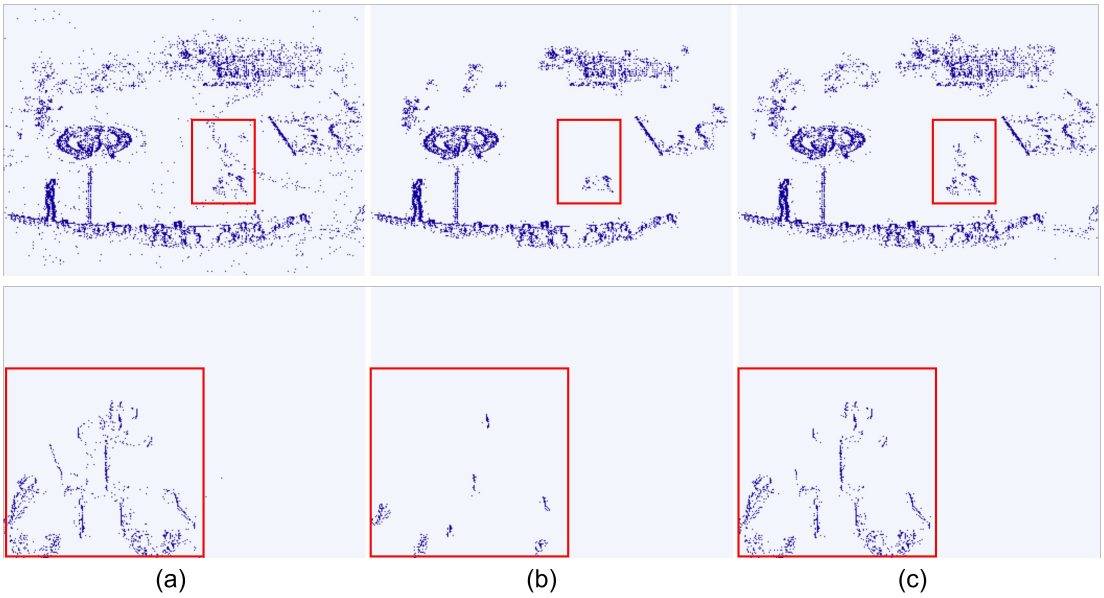


Fig. 6: Visual comparison of event processing results. (a) Raw event stream input; (b) Events filtered by classical LIF neurons with fixed parameters; (c) Events processed by our proposed ASN module. The ASN module effectively suppresses scattered noise while preserving dense event clusters corresponding to moving UAV/vehicle.

module enables the network to focus more quickly on moving targets and reduces the number of training iterations required for convergence.

Table 3: Comparison of different fusion strategies on the FE108 dataset. Best results are highlighted in bold.

Strategy	**Add**	**Concat**	**CMT**[16]	**SREF**[45]	**CDMS**[15]	**MCP**[58]	**HSTF(Ours)**
RPR (%)	62.4	86.0	88.4	87.9	87.5	87.9	**89.0**
RSR (%)	40.3	56.7	57.8	57.4	**58.7**	56.9	57.8

HSTF module: The impact of our High-order Spatio-Temporal Feature fusion module is evaluated by comparing **Variant E** (full STC-Net) against **Variant D** (ASN only). Variant E achieves the best performance with 89.0% RPR and 57.8% RSR, showing additional improvements of +3.0% RPR and +1.1% RSR. These gains demonstrate that simple concatenation fusion (Variant D) is insufficient for fully exploiting RGB-Event complementarity. Table 3 shows the results of different fusion strategies, where our HSTF module outperforms all compared methods. Notably, HSTF achieves 89.0% RPR, surpassing CMT (88.4%), SREF (87.9%), CDMS (87.5%), and MCP (87.9%), while maintaining competitive RSR performance. These results validate the effectiveness of our high-order interaction design. Our HSTF module enables deeper synergistic integration through: (1) *Frequency-domain interactions* that efficiently capture global spatial correlations while avoiding computational overhead; (2) *High-order spatial interactions* via multi-context modulation that selectively emphasize informative regions across modalities; (3) *High-order channel interactions* that model complex inter-dependencies through recursive squeeze-and-excitation mechanisms. By progressively integrating information across frequency, spatial, and channel domains, this hierarchical fusion strategy ensures comprehensive multi-dimensional feature representation.

The complete improvement trajectory from Variant A (74.7% RPR) to Variant E (89.0% RPR) reveals strong synergistic effects among all components. The total gain of 14.3% RPR

demonstrates that: (1) Event data provides crucial temporal information missing in RGB; (2) ASN processing optimizes event representations for tracking; (3) HSTF fusion maximizes cross-modal complementarity. Importantly, the interaction between ASN and HSTF is multiplicative rather than additive—ASN provides cleaner event features that HSTF can more effectively fuse with RGB semantics. This synergy explains why our full STC-Net substantially outperforms both unimodal approaches and simple fusion baselines.

5 Conclusion

In this paper, we propose a Spatio-Temporal Collaboration Network (STC-Net) for robust RGB-Event object tracking, which consists of two key components: an Adaptive Spiking Neuron (ASN) module and a High-order Spatio-Temporal Feature (HSTF) fusion module. The proposed ASN module effectively filters noise and enhances motion-related events through adaptive threshold and membrane dynamics. Subsequently, the HSTF module enables deep cross-modal interaction by exploring high-order synergistic relationships in frequency, spatial, and channel dimensions. Extensive experiments on the FE108 and VisEvent benchmarks demonstrate that our STC-Net achieves state-of-the-art performance in various challenging conditions. The proposed method offers an effective approach to leveraging the complementary strengths of frame and event modalities for reliable object tracking.

Declarations

Conflict of interest. The authors declare no conflict of interest.

Data Availability. No datasets were generated or analysed during the current study.

Author contributions. Nan Wang designed the research, conceived the experiments, conducted experimental demonstrations, and wrote the manuscript. Yong Song, Li Wang, and Yingbo He provided economic and technical support. Shaoxing Wu and Teng Luo contributed to the algorithm design. Shuang Wang and Linjun Zeng contributed to the data collection and analysis. Yong Song, Shaoxing Wu, and Ya Zhou commented on the manuscript. All authors read and approved the final manuscript.

Funding. This work was supported by National Natural Science Foundation of China General Program (82272130) and National Natural Science Foundation of China Key Program (U22A20103).

References

[1] Zhu, J., Zhang, Q., Fei, L., Cai, R., Xie, Y., Sheng, B., Yang, X.: Fffn: Frame-by-frame feedback fusion network for video super-resolution. IEEE Transactions on Multimedia **25**, 6821–6835 (2022)

[2] Lin, C., Zou, C., Xu, H.: Scnet: A dual-branch network for strong noisy image denoising based on swin transformer and convnext. Computer Animation and Virtual Worlds **36**(3), 70030 (2025)

[3] Cui, J., Pang, L., Zhao, S., Peng, Z., Hu, X., Zeng, L., Jiang, T., Liang, M., Huang, J., Yuan, W., *et al.*: Learning pulse image with deep dynamic frequency network for cardiovascular diseases diagnosis: Ji cui et al. The Visual Computer **41**(12), 9717–9735 (2025)

[4] Jiang, M., Tian, Z., Yu, C., Shi, Y., Liu, L., Peng, T., Hu, X., Yu, F.: Intelligent 3d garment system of the human body based on deep spiking neural network. Virtual Reality & Intelligent Hardware **6**(1), 43–55 (2024)

[5] Wu, Q., Yan, Y., Liang, Y., Liu, Y., Wang, H.: Dsnet: Deep and shallow feature learning for efficient visual tracking. In: Asian Conference on Computer Vision, pp. 119–134. Springer, Cham (2018)

[6] Chen, B., Li, P., Sun, C., Wang, D., Yang, G., Lu, H.: Multi attention module for visual tracking. Pattern Recognition **87**, 80–93 (2019)

[7] Bhat, G., Johnander, J., Danelljan, M., Khan, F.S., Felsberg, M.: Unveiling the power of deep tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–498 (2018)

[8] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision, pp. 850–865. Springer, Cham (2016)

[9] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8126–8135 (2021)

[10] Brandli, C., Berner, R., Yang, M., Liu, S.-C., Delbruck, T.: A 240×180 130db 3μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014)

[11] Gallego, G., Delbruck, T., Orchard, G.M., Bartolozzi, C., Scaramuzza, D.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2020)

[12] Gehrig, D., Rebecq, H., Gallego, G., Scaramuzza, D.: Asynchronous, photometric feature tracking using events and frames. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 750–765 (2018)

[13] Yang, Z., Wu, Y., Wang, G., Yang, Y., Li, G., Deng, L., Zhu, J., Shi, L.: Dashnet: A hybrid artificial and spiking neural network for high-speed object tracking. arXiv preprint arXiv:1909.12942 (2019)

[14] Liu, H., Moeys, D.P., Das, G., Neil, D., Liu, S.-C., Delbrück, T.: Combined frame-and event-based detection and tracking. In: 2016 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2511–2514. IEEE, New York, NY (2016)

[15] Zhang, J., Yang, X., Fu, Y., Wei, X., Yin, B., Dong, B.: Object tracking by jointly exploiting frame and event domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13043–13052 (2021)

[16] Wang, X., Li, J., Zhu, L., Zhang, Z., Chen, Z., Li, X., Wang, Y., Tian, Y., Wu, F.: Visevent: Reliable object tracking via collaboration of frame and event flows. IEEE Transactions on Cybernetics **54**(3), 1997–2010 (2023)

[17] Zhang, J., Wang, Y., Liu, W., Li, M., Bai, J., Yin, B., Yang, X.: Frame-event alignment and fusion network for high frame rate tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9781–9790 (2023)

[18] Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, pp. 2544–2550 (2010). IEEE

[19] Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized

correlation filters. IEEE transactions on pattern analysis and machine intelligence **37**(3), 583–596 (2014)

[20] Wang, M., Liu, Y., Huang, Z.: Large margin object tracking with circulant feature maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4021–4029 (2017)

[21] Yang, T., Chan, A.B.: Recurrent filter learning for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2010–2019 (2017)

[22] Huang, Y., Liu, S., Dong, F., Li, X., Yang, X., Zhou, Y., Huang, J., Song, Y.: Pl-mct: pseudo-labeling and multi-frame consistency training for semi-supervised visual tracking. The Visual Computer **41**(6), 4143–4156 (2025)

[23] Marvasti-Zadeh, S.M., Cheng, L., Ghanei-Yakhdan, H., Kasaei, S.: Deep learning for visual tracking: A comprehensive survey. IEEE Transactions on Intelligent Transportation Systems **23**(5), 3943–3968 (2021)

[24] Wen, Y., Luo, B., Shi, W., Ji, J., Cao, W., Yang, X., Sheng, B.: Sat-net: structure-aware transformer-based attention fusion network for low-quality retinal fundus images enhancement. IEEE Transactions on Multimedia (2025)

[25] Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120db 15μs latency asynchronous temporal contrast vision sensor. IEEE journal of solid-state circuits **43**(2), 566–576 (2008)

[26] Cai, L., Zhou, Z.: Visual expansion and real-time calibration for pan-tilt-zoom cameras assisted by panoramic models. Computer Animation and Virtual Worlds **36**(1), 70015 (2025)

[27] Litzenberger, M., Posch, C., Bauer, D., Belbachir, A.N., Schon, P., Kohn, B., Garn, H.: Embedded vision system for real-time object tracking using an asynchronous transient vision sensor. In: 2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop, pp. 173–178. IEEE, New York, NY (2006)

[28] Pikatkowska, E., Belbachir, A.N., Schraml, S., Gelautz, M.: Spatiotemporal multiple persons tracking using dynamic vision sensor. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 35–40 (2012). IEEE

[29] Chen, H., Suter, D., Wu, Q., Wang, H.: End-to-end learning of object motion estimation from retinal events for event-based object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10534–10541 (2020)

[30] Yao, M., Gao, H., Zhao, G., Wang, D., Lin, Y., Yang, Z., Li, G.: Temporal-wise attention spiking neural networks for event streams classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10221–10230 (2021)

[31] Gerstner, W., Kistler, W.M., Naud, R., Paninski, L.: Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition. Cambridge University Press, Cambridge, UK (2014)

[32] Wu, S., Wang, G., Song, Y., Zhao, Y., Zhou, Y., Meng, Q., Liao, Y.: Bidirectional dynamic threshold snn for enhanced object detection with rich spike information. Frontiers in Neuroscience **19**, 1661916 (2025)

[33] Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., Tian, Y.: Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2661–2671 (2021)

[34] Zhang, J., Dong, B., Zhang, H., Ding, J., Heide, F., Yin, B., Yang, X.: Spiking transformers for event-based single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8801–8810 (2022)

[35] Nagaraj, M., Liyanagedera, C.M., Roy, K.: Dotie-detecting objects through temporal isolation of events using a spiking architecture. arXiv preprint arXiv:2210.00975 (2022)

[36] Zhang, J., Zhang, M., Wang, Y., Liu, Q., Yin, B., Li, H., Yang, X.: Spiking neural networks with adaptive membrane time constant for event-based tracking. IEEE Transactions on Image Processing (2025)

[37] Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., Huang, L.: What makes multi-modal learning better than single (provably). Advances in Neural Information Processing Systems **34**, 10944–10956 (2021)

[38] Yang, J., Gao, S., Li, Z., Zheng, F., Leonardis, A.: Resource-efficient rgbd aerial tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13374–13383 (2023)

[39] Li, C., Cheng, H., Hu, S., Liu, X., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25**(99), 5743–5756 (2016)

[40] Li, C., Zhao, N., Lu, Y., Zhu, C., Tang, J.: Weighted sparse representation regularized graph learning for rgb-t object tracking. In: Acm on Multimedia Conference, New York, NY, pp. 1856–1864 (2017)

[41] Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: Benchmark and baseline. Pattern Recognition **96**, 106977 (2019)

[42] Xiao, J., Stolkin, R., Gao, Y., Leonardis, A.: Robust fusion of color and depth data for rgb-d target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. IEEE transactions on cybernetics **48**(8), 2485–2499 (2017)

[43] Zhang, P., Zhao, J., Bo, C., Wang, D., Lu, H., Yang, X.: Jointly modeling motion and appearance cues for robust rgb-t tracking. IEEE Transactions on Image Processing **30**, 3335–3347 (2021)

[44] Luo, C., Sun, B., Yang, K., Lu, T., Yeh, W.-C.: Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme. Infrared Physics & Technology **99**, 265–276 (2019)

[45] Cao, J., Zheng, X., Lyu, Y., Wang, J., Xu, R., Wang, L.: 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), pp. 9026–9032 (2024). IEEE

[46] Liu, S., Wang, G., Song, Y., Huang, J., Huang, Y., Zhou, Y., Wang, S.: Siameft: adaptive-time feature extraction hybrid network for rgbe multi-domain object tracking. Frontiers in Neuroscience **18**, 1453419 (2024)

[47] Zheng, N., Zhou, M., Huang, J., Hou, J., Li, H., Xu, Y., Zhao, F.: Probing synergistic high-order interaction in infrared and visible image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26384–26395 (2024)

[48] Sun, H., Liu, R., Cai, W., Wang, J., Wang, Y., Tang, H., Cui, Y., Yao, D., Guo, D.: Reliable object tracking by multimodal hybrid feature extraction and transformer-based fusion.

Neural Networks **178**, 106493 (2024)

[49] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[50] Zhang, L., Danelljan, M., Gonzalez-Garcia, A., Van De Weijer, J., Shahbaz Khan, F.: Multi-modal fusion for end-to-end rgb-t tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)

[51] Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6182–6191 (2019)

[52] Wen, Y., Shen, S., Shi, W., Cao, W., Bi, L., Yang, X., Sheng, B.: A lightweight depthwise separable convnet with frequency-domain enhancement for retinal vessel segmentation. ACM Transactions on Multimedia Computing, Communications and Applications **21**(12), 1–23 (2025)

[53] Dong, X., Shen, J., Shao, L., Porikli, F.: Clnet: A compact latent network for fast adjusting siamese trackers. In: European Conference on Computer Vision, Cham, pp. 378–395 (2020). Springer

[54] Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7183–7192 (2020)

[55] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4282–4291 (2019)

[56] Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4660–4669 (2019)

[57] Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G.: Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12549–12556 (2020)

[58] Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9516–9526 (2023)