EvRT-DETR: Latent Space Adaptation of Image Detectors for Event-based Vision

Dmitrii Torbunov[1], Yihui Ren[1], Animesh Ghose[1], Odera Dim[1], Yonggang Cui[1]
[1]Brookhaven National Laboratory, Upton, NY, USA

dtorbunov@bnl.gov, yren@bnl.gov, aghose@bnl.gov, dodera@bnl.gov, ycui@bnl.gov

Abstract

Event-based cameras (EBCs) have emerged as a bio-inspired alternative to traditional cameras, offering advantages in power efficiency, temporal resolution, and high dynamic range. However, development of image analysis methods for EBCs is challenging due to the sparse and asynchronous nature of the data. This work addresses the problem of object detection for EBC cameras. The current approaches to EBC object detection focus on constructing complex data representations and rely on specialized architectures. We introduce I2EvDet (Image-to-Event Detection), a novel adaptation framework that bridges mainstream object detection with temporal event data processing. First, we demonstrate that a Real-Time DEtection TRansformer, or RT-DETR, a state-of-the-art natural image detector, trained on a simple image-like representation of the EBC data achieves performance comparable to specialized EBC methods. Next, as part of our framework, we develop an efficient adaptation technique that transforms image-based detectors into event-based detection models by modifying their frozen latent representation space via minimal architectural additions. The resulting EvRT-DETR model reaches state-of-the-art performance on the standard benchmark datasets Gen1 (mAP $+2.3$*) and 1Mpx/Gen4 (mAP* $+1.4$*). These results demonstrate a fundamentally new approach to EBC object detection through principled adaptation of mainstream architectures, offering an efficient alternative with potential applications to other temporal visual domains. The code is available at:* https://github.com/realtime-intelligence/evrt-detr.

1. Introduction

Event-based cameras (EBCs) present a biologically inspired alternative to traditional frame-based cameras. Unlike conventional cameras that capture frames at predetermined intervals, EBC pixels operate asynchronously, generating data only when brightness changes exceed a threshold. This results in sparse, asynchronous data streams where each event is a tuple containing pixel location, timestamp, and polarity (whether brightness increased or decreased), as illustrated in Figure 2. This approach offers remarkable advantages: low power consumption (as low as 10 mW), reduced data transfer rates, exceptional temporal resolution (on the order of μs), and high dynamic range (>100 dB). These properties have led to widespread adoption in autonomous driving, robotics, and wearable electronics [6].

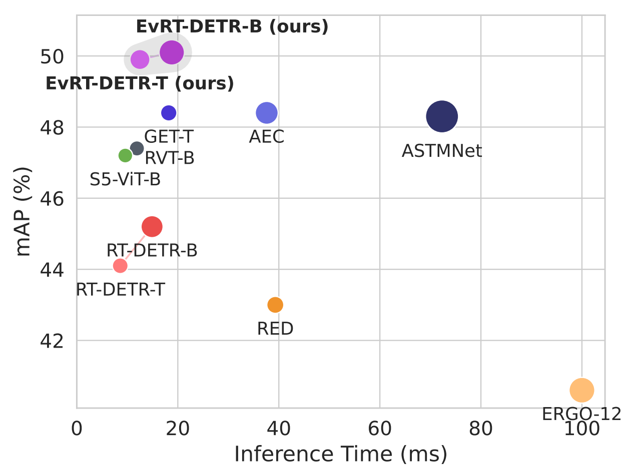

Figure 1. **Object Detection Performance vs. Inference Time.** A summary of object detection performance (COCO mAP) versus inference time (ms) of various models on the 1Mpx automotive dataset. Circle size is proportional to the number of model parameters. Inference times are reported for NVIDIA T4 GPU. Our EvRT-DETR models achieve state-of-the-art accuracy while maintaining competitive inference speeds.

These unique characteristics of EBC data – sparsity, asynchronicity, and temporal nature – create fundamental challenges when applying traditional computer vision techniques. Early attempts to directly apply conventional

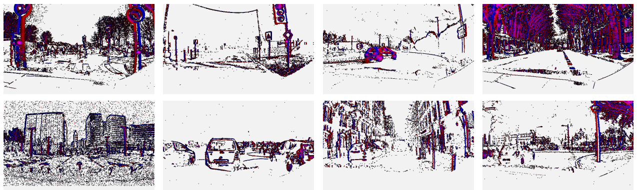

Figure 2. **Event-based Camera Output Visualization.** Sample frames from the 1Mpx dataset showing event polarity. Red indicates positive events, and blue denotes negative events.

image-based object detectors to simple event representations yielded poor performance [22], highlighting the gap between these domains. This led to two predominant research directions: (1) constructing sophisticated event representations that transform EBC data into formats compatible with existing detectors (e.g., [33]) or (2) designing specialized architectures tailored to handle the temporal nature of event data (e.g., [20]). A common assumption in these approaches is that EBC data processing requires either specialized representations or architectures due to the fundamental differences between event- and frame-based vision, leading to development paths that often diverge from mainstream computer vision research.

Meanwhile, mainstream object detection has seen remarkable progress, evolving from convolutional neural network (CNN)-based architectures, e.g., YOLO [24] and Faster R-CNN [25], to transformer-based models like DETR (DEtection TRansformer) [2]. Recent advances have culminated in highly efficient and accurate models, such as the **R**eal-**T**ime **DE**tection **TR**ansformer (RT-DETR) [30] that achieves exceptional performance while maintaining real-time inference speeds. Though primarily optimized for processing individual images, these state-of-the-art detectors offer powerful feature extraction capabilities that potentially could be leveraged for temporal data.

This evolution in mainstream detection creates an opportunity to bridge the gap between conventional and event-based vision. Rather than assuming fundamental incompatibility between these domains, we explore a complementary direction: adapting state-of-the-art frame-based detectors to process event data effectively with minimal modifications. In this work, we investigate how RT-DETR can be strategically adapted to the event-based domain, demonstrating that modern object detection architectures can successfully work on event data and opening a new research direction.

Our approach proceeds in two stages to systematically adapt RT-DETR to event data. First, we train RT-DETR directly on simple image-like representations of EBC data using standard training procedures without special modifications. This straightforward approach yields surprising results – even when processing only fixed $50\,\mathrm{ms}$ time frames, the detector achieves performance comparable to specialized EBC-specific methods explicitly designed for video object detection on EBC data. This is particularly unexpected as these specialized methods deliberately exploit the temporal nature of EBC data and process information from the distant past, while our approach performs detection using only a single fixed time window (frame) of $50\,\mathrm{ms}$, similar to traditional frame-based cameras.

Building on this foundation, we introduce **I2EvDet (Image-to-Event Detection)**, a technique that enables straightforward transformation of image-based object detectors into video-capable models. This approach leverages latent space adaptation principles [14, 15] where we freeze the pre-trained RT-DETR model and insert a lightweight recurrent neural network (RNN) module within the encoder's latent representation space. This design allows the model to capture temporal dynamics across frames while preserving the powerful representation capabilities of the original detector and adding minimal computational overhead. The resulting adaptation effectively leverages information across multiple time windows without requiring a specialized video detection architecture.

The combined approach, called **EvRT-DETR**, achieves state-of-the-art performance on standard EBC benchmarks, outperforming specialized methods on both Gen1 [4] (mAP $+2.3$) and 1Mpx [22] (mAP $+1.4$) datasets. Importantly, the results are achieved using only standard modules from natural image and video analysis *without* requiring domain-specific modifications. This demonstrates not only the effectiveness of our approach for EBC data, but also suggests a broader paradigm for adapting image detectors to temporal domains. As shown in Figure 1, our model achieves superior accuracy while maintaining competitive inference speeds, making it practical for real-world applications.

Contributions of this work:

- We show that when trained on simple image-like representations of EBC data using standard procedures, mainstream object detectors like RT-DETR can achieve performance comparable to specialized EBC methods, offering a complementary approach to domain-specific architectures.
- We propose I2EvDet, an efficient adaptation technique that transforms image-based detectors into video-capable models through minimal architectural additions to the frozen pre-trained model and with potential applications to various temporal visual domains.
- We validate our approach in the challenging domain of event-based cameras, where EvRT-DETR achieves state-of-the-art performance on the standard Gen1 (mAP $+2.3$) and 1Mpx (mAP $+1.4$) EBC benchmarks while maintaining competitive inference speeds.

2. Related Work

Research at the intersection of event-based cameras and object detection spans multiple domains in computer vision. Current approaches primarily focus on two directions: designing efficient representations for the unique properties of event data and developing specialized neural architectures to process these representations. However, despite its success in other domains, a third direction – adapting mainstream computer vision models to event data – remains underexplored. This section reviews relevant work across these areas, highlighting how our approach bridges the gap between specialized event-based methods and mainstream object detection advances.

2.1. EBC Data Representations

An event camera produces a stream of event data in the form (t, p, x, y), where t is the timestamp of an event, p is the polarity (positive or negative) of the brightness change, and (x, y) is the location of the event pixel. To simplify the analysis, the event data often are transformed into alternative representations more suitable for conventional image analysis algorithms.

The simplest image-like representations of event camera data are *Event Frames* or *2D Histograms* [6], where events within fixed time windows are accumulated into two-dimensional frames of shape $(2, H, W)$ with polarity as the first dimension and spatial coordinates as the remaining dimensions. A natural evolution is the *Stacked 2D Histogram*, which further partitions each frame into T time intervals, creating a representation of shape $(2, T, H, W)$ that typically is reshaped to $(2T, H, W)$ and treated as a natural image. While these representations are straightforward, directly applying existing computer vision algorithms to them has yielded poor performance in the past [22].

To improve performance, other representations are being explored. For example, a *Time Surface (TS)* representation [6, 33] is an image-like representation where pixel values encode the time elapsed since the last event in a given pixel. Thus, unlike the simplest fixed time window representations, *TS* potentially can encode arbitrarily distant past. Empirically, it has been shown that using the *TS* representations offers significantly better performance on the object detection tasks compared to the *2D Histogram* representations [22].

Many other EBC data representations also have been proposed. The recent ERGO-12 (Event Representation through Gromov-Wasserstein Optimization) [33] work has developed efficient image-like representations, achieving state-of-the-art object detection performance on the Gen1 dataset. Meanwhile, other approaches reformulate object detection as a three-dimensional (3D) problem by treating events as points in spatial-temporal space or discretizing them into voxel grids [6]. While computationally heavier, these representations can preserve exact event timing, potentially enabling better handling of overlapping objects and complex motions.

2.2. EBC Object Detection Architectures

Once the data representation is selected, one can start to develop object detection algorithms. There are several families of object detection methods designed for particular data representations.

The most straightforward approach applies conventional image detection methods directly to *2D Histogram* frames, but this yields relatively poor performance [22]. More advanced methods, such as Recurrent Vision Transformer (RVT) [9] and State Space Vision Transformer (S5-ViT-B) [34], extend this approach by incorporating temporal memory to capture sequential information, significantly improving detection quality compared to frame-by-frame processing.

The most recent family of methods attempts to use more efficient event representations to further improve object detection performance. For example, 12-channel ERGO-12 [33] constructs an optimized image-like event representation and shows that the standard object detection methods perform exceptionally well on it. Alternatively, Asynchronous Spatio-Temporal Memory Network (ASTMNet) [16] and Adaptive Event Conversion (AEC) [19] attempt to combine better event representation while using novel neural architectures to expand state-of-the-art results.

Other methods have been developed that treat EBC object detection as a 3D detection problem [8, 22, 26], exploit benefits of neuromorphic architectures [3], or explore various hybrid approaches [20, 21].

2.3. Transformer-based Object Detection

Object detection has been revolutionized by transformer architectures in recent years. DETR [2] was the first widely successful transformer-based [28] model for object detection, offering a simple and elegant alternative to traditional CNN-based architectures [10, 24].

DETR follows a hybrid CNN-Transformer design with a backbone-encoder-decoder structure. It uses a CNN backbone (typically ResNet [13]) for feature extraction, followed by a Transformer encoder that processes these features into tokens and captures their correlations. The Transformer decoder then identifies objects through cross-attention to the encoder outputs.

Despite its innovative approach, the original DETR suffered from limitations, including training instability [29], poor performance on small objects [2], and high computational costs during inference [32]. Subsequent research has addressed these issues through architectural improvements such as Feature Pyramid Networks (FPN), specialized decoder tokens, and deformable attention mechanisms [2, 32].

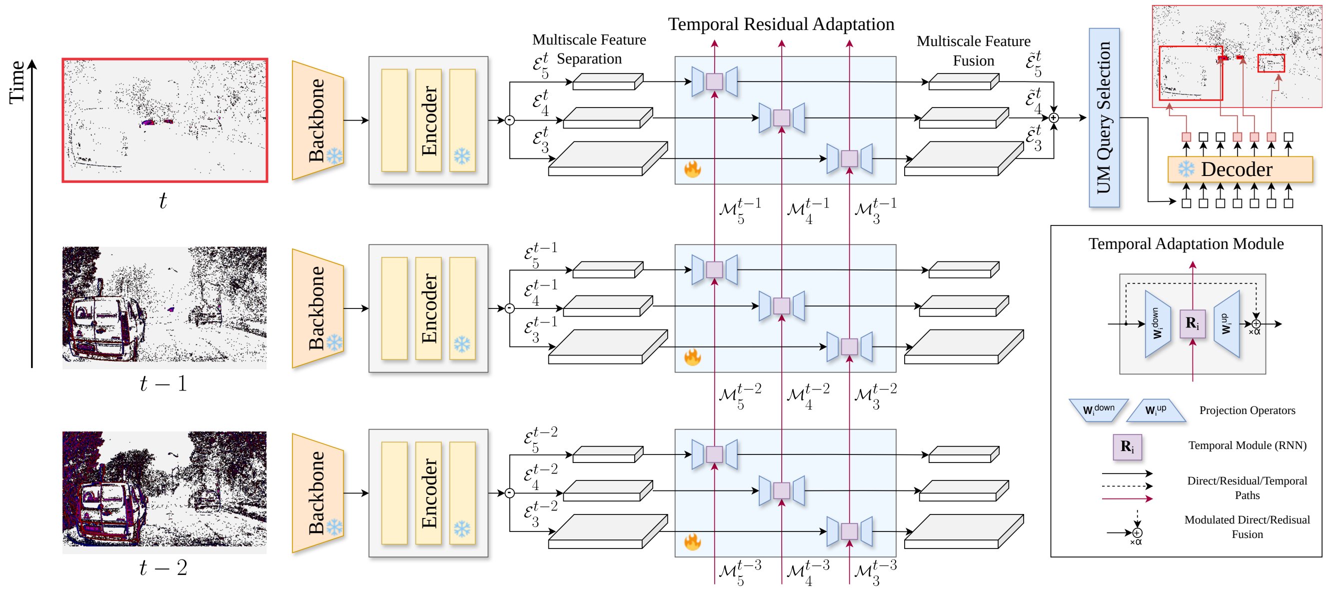

Figure 3. **Overview of the I2EvDet Framework Applied to RT-DETR**. This approach adapts a pre-trained RT-DETR model to process temporal event data using minimal architectural modifications. The backbone and encoder components remain frozen during temporal adaptation, while trainable temporal modules are strategically inserted at multiple feature scales. Left: Event camera frames are processed through the frozen RT-DETR backbone and encoder. Center: Our temporal adaptation modules operate on multiscale features through residual connections (red pathways), enabling information flow across time steps while preserving spatial representation integrity. Right: Feature fusion and decoder stages produce final detections. The detailed view of the Temporal Adaptation Module (bottom right) shows the projection operations (W_i^{down}, W_i^{up}) and recurrent memory module ($\mathbf{R}_i$) that enable efficient temporal modeling with minimal additional parameters.

These incremental advances have culminated in the RT-DETR [30], which achieves exceptional performance across all object scales with stable training and significantly faster inference times. RT-DETR represents the current state-of-the-art in efficient transformer-based detection and serves as an ideal foundation for adaptation to specialized domains, such as event-based vision.

2.4. Model Adaptation Techniques

While specialized architectures dominate event-based vision, adapting mainstream architectures offers compelling advantages. Parameter-efficient methods [12, 14, 15] modify pre-trained models with minimal changes by adding compact modules to frozen backbones. These approaches have proven effective across language [5], vision [31], and multimodal systems [11], but they remain underexplored for event-based vision despite their efficiency and transfer learning benefits.

Given the success of adaptation techniques in other domains and the power of transformer-based detection models, there is a clear opportunity to bridge these approaches for event-based vision. Our work explores this intersection, demonstrating that mainstream object detectors can be effectively adapted to EBC data with minimal modifications rather than requiring specialized architectures.

3. Methods

This section presents I2EvDet, our adaptation framework for event-based object detection. Unlike approaches that rely on specialized architectures or complex representations, I2EvDet bridges mainstream object detection with temporal event data via strategic, minimal modifications. Our framework demonstrates how image-based detectors can be efficiently transformed into video-capable models while preserving their core strengths. We first describe our image-like representation of event data then introduce the I2EvDet framework. Finally, we detail the technical implementation and discuss how our approach addresses the unique temporal dependencies in event camera data.

3.1. Event Representation

Our adaptation framework begins with transforming EBC data into an image-like representation that is compatible with mainstream object detection models. We employ the *Stacked 2D Histogram* representation [6] as it provides a straightforward bridge between event data and conventional image processing architectures.

To ensure direct comparability with existing methods, we adopt the frame construction parameters from RVT [9], which allows us to isolate the impact of our architectural choices from data pre-processing. Specifically, we take the stream of events and partition it into a series of consecutive fixed time windows of $T_{\text{frame}} = 50\,\text{ms}$. Each such window corresponds to a single frame. Next, we further subdivide each frame into 10 intervals of $T_{\text{bin}} = 5\,\text{ms}$.

To construct a frame corresponding to an interval $[t_0, t_0 + T_{\text{frame}}]$, we create an intermediate stacked histogram S from a set of events $\mathcal{E}$ in that interval:

$$S(p, t_i, y, x) = \sum_{\{e \mid t_e \in [t_0, t_0 + T_{\text{frame}})\}} \delta_x^{x_e} \delta_y^{y_e} \delta_p^{p_e} \mathbf{1}_{[t_i, t_{i+1})}(t_e)$$

where $e = (t_e, p_e, x_e, y_e)$ represents an individual event with time stamp t_e, polarity p_e, and spatial coordinates (x_e, y_e); $t_i = t_0 + i \cdot T_{\text{bin}}$, $i \in [0, 10)$ is the histogram bin index; (x, y) are the spatial indices in the output histogram; δ_a^b is the Kronecker delta function; and $\mathbf{1}$ is an indicator function. Once a stacked histogram S of shape $(2, 10, H, W)$ is constructed, we merge the polarity and time dimensions to obtain an image-like 2D frame F of shape $(20, H, W)$.

3.2. I2EvDet: Image-to-Event Detection Framework

I2EvDet is a principled framework for adapting image-based object detectors to process temporal data. While specialized architectures typically are developed for event-based object detection, this approach demonstrates that strategic adaptation of mainstream detectors can achieve superior performance with minimal architectural modifications.

The I2EvDet framework applies latent space adaptation principles [12, 14, 15], which enable domain adaptation through minimal changes to pre-trained models. The I2EvDet framework follows a two-stage process, and it is applicable to any object detection model with a clean separation between feature extraction and object detection modules (e.g., YOLO and DETR families).

First, we establish a solid foundation by training an object detector on individual EBC frames (subsection 3.1), treating them as conventional images. This creates a detector with powerful spatial representation capabilities optimized for event data.

In the second stage, we transform this image-based detector into a video-capable model. We freeze the pre-trained detector parameters and insert lightweight RNN modules between the feature extractor and detection components. We train the inserted temporal modules on the EBC video object detection task while keeping the original detector frozen. The resulting model captures temporal dynamics across frames while preserving the powerful representation capabilities of the original detector and adding minimal computational overhead.

This two-stage approach retains learned spatial representations of the original image detector while efficiently adding temporal capabilities. In this work, we apply I2EvDet to the RT-DETR model as the primary detector, resulting in the EvRT-DETR model. Additionally, we demonstrate the generalizability of the I2EvDet framework using experiments with YOLOX detectors.

3.3. Temporal Modeling Considerations

A key theoretical consideration for this adaptation framework is the choice of temporal memory mechanism. Event cameras present a unique challenge: static objects become "invisible" as they generate no events when motionless. Their detection requires having a persistent memory of past events. While transformers have dominated recent vision architectures, their fixed context windows limit temporal memory. We employ RNNs for their theoretically unbounded memory capacity through recurrent state, making them uniquely suited for maintaining object persistence during periods of invisibility in the event stream. This theoretical foundation has informed the EvRT-DETR implementation.

3.4. Technical Implementation of EvRT-DETR

The EvRT-DETR architecture implements the I2EvDet framework via strategic, minimal modifications to the RT-DETR model. As shown in Figure 3, we identify key integration points within the pre-trained model's latent space, where temporal information can be effectively incorporated without disrupting spatial feature extraction.

The RT-DETR encoder produces multiscale feature representations $\{\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5\}$, where each element corresponds to a transformer token encoding rich spatial information. These features, originating from the last three stages of the backbone $\{\mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\}$ [30], provide an ideal latent manifold for temporal adaptation.

Following latent space adaptation principles, we introduce specialized temporal processing modules at each scale, implementing three parallel RNN-based memory modules $\{\mathbf{R}_3, \mathbf{R}_4, \mathbf{R}_5\}$. This multiscale temporal adaptation design ensures that temporal dynamics are captured at different levels of feature abstraction. Critically, each RNN module interfaces with its corresponding encoder feature map through residual connections, mathematically expressed as:

$$\mathcal{E}_i^{t,proj} = W_i^{down} \cdot \mathcal{E}_i^t \tag{1}$$

$$\left(\mathcal{O}_i^{t,proj}, \mathcal{M}_i^t\right) = \mathbf{R}_i \left(\mathcal{E}_i^{t,proj}, \mathcal{M}_i^{t-1}\right) \tag{2}$$

$$\mathcal{O}_i^t = W_i^{up} \cdot \mathcal{O}_i^{t,proj} \tag{3}$$

$$\tilde{\mathcal{E}}_i^t = \mathcal{E}_i^t + \alpha_i \cdot \mathcal{O}_i^t, \tag{4}$$

where $W_i^{down} \in \mathbb{R}^{C_i^{RNN} \times C_i}$ and $W_i^{up} \in \mathbb{R}^{C_i \times C_i^{RNN}}$ are the projection matrices that transform features into and out of the temporal processing space, $\mathcal{M}_i^t \in \mathbb{R}^{C_i^{RNN} \times H_i \times W_i}$

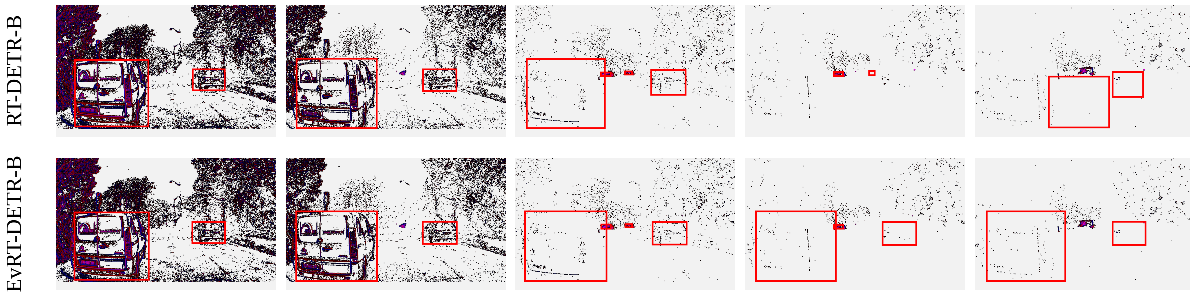

Figure 4. **Temporal Adaptation Benefits of EvRT-DETR.** Visualization of object detection across sequential 1Mpx frames captured by a vehicle-mounted camera that gradually comes to a stop. While the camera is in motion (left frames), both detectors maintain accurate bounding boxes of stationary vehicles. As camera motion decreases (middle frames), event data becomes increasingly sparse. When the camera stops completely (right frames), RT-DETR (top row) loses detection capability, while EvRT-DETR (bottom row) maintains consistent detection by engaging its temporal memory module.

represents the temporal memory state for the i-th feature scale at time t, $\mathcal{O}_i^t$ denotes the corresponding temporal adaptation output, and α_i is a learnable scaling factor that modulates the influence of temporal information on the spatial representation inspired by the ReZero technique [1]. In our implementation, the input projection matrix W_i^{down} is integrated within the RNN architecture, which already handles dimensionality reduction through its internal gating mechanisms. Meanwhile, the output projection W_i^{up} is explicitly implemented to ensure proper integration back into the frozen RT-DETR architecture. Thus, the transformed feature representation $\tilde{\mathcal{E}}_i^t$ incorporates temporal context while maintaining the integrity of the original spatial representation $\mathcal{E}_i^t$. In addition, our ablation studies show that C_i^{RNN} can be reduced 4x relative to C_i with minimal performance impact.

For specific implementation of the RNN modules, we employ the ConvLSTM architecture [27], which effectively balances representational capacity with computational efficiency. This choice was informed by ConvLSTM's demonstrated effectiveness in handling spatio-temporal data in the RVT model [9], allowing our adaptation to employ established temporal modeling techniques.

4. Evaluation of the EvRT-DETR Model

The EvRT-DETR model is evaluated on two standard EBC object detection benchmarks: Gen1 [4] and 1Mpx [22] (also known as Gen4) datasets. These datasets contain diverse driving scenarios captured by Prophesee's neuromorphic vision sensors, providing an ideal testbed for validating our adaptation approach under real-world conditions.

4.1. Experimental Setting

Datasets. We evaluate on two standard Prophesee automotive datasets: Gen1 [4] and 1Mpx [22]. Gen1 contains 39+ hours of recordings at 304×240 resolution with manual annotations (1-4 Hz) for cars and pedestrians, presenting challenges with sparse labels and stationary objects. 1Mpx offers 14+ hours at higher resolution (1280×720) with denser automatic annotations (60 Hz) for cars, pedestrians, and two-wheelers. Both datasets are organized as 60-second clips with standard train/test splits. 1Mpx's higher resolution and frame rate enable evaluation on more complex scenes, while Gen1's annotation pattern tests robustness to sparse temporal supervision.

Event Representations. We adopt the well-established *stacked 2D Histogram* approach from RVT protocols [9], constructing event frames from each $50\,\mathrm{ms}$ time window. Each frame is subdivided into 10 equally time-spaced $5\,\mathrm{ms}$ intervals, creating a representation with sufficient temporal resolution for our adaptation framework while remaining compatible with standard image processing architectures.

For the Gen1 dataset (original size $(240, 304)$), we pad frames to $(256, 320)$ to ensure divisibility by 32 as required by the feature pyramid architecture of RT-DETR. For 1Mpx, we downsize from $(720, 1280)$ to $(360, 640)$ using bilinear interpolation and pad to $(384, 640)$. This maintains consistency with prior work while allowing for fair computational comparison.

Two-stage Training Process. We train EvRT-DETR in two stages: (1) training the RT-DETR detector on individual event frames using standard practices (Adam optimizer; EMA of weights) then (2) freezing these parameters and training only the lightweight RNN modules on both random and sequential clips to develop temporal memory. The first stage establishes robust spatial feature extraction, while the second adds temporal processing with minimal parameter overhead. For temporal training, we maintain RNN memory continuity across sequential clips while resetting for random clips using different clip lengths for Gen1 (21 frames) and 1Mpx (10 frames). Full implementation details are provided in the supplementary material.

Model	Gen1		1Mpx/Gen4		Params (M)
	mAP (%)	Runtime (ms)	mAP (%)	Runtime (ms)	
AsyNet [18]	14.5	-	-	-	11.4
AEGNN [26]	16.3	-	-	-	20.0
Spiking DenseNet [3]	18.9	-	-	-	8.2
Events-RetinaNet [22]	34.0	-	18.0	-	32.8
RED [22]	40.0	16.7	43.0	39.3	24.1
AEC [19]	47.0	10.6 (3.9)	48.4	37.6 (13.8)	46.5
RVT-B [9]	47.2	10.2	47.4	11.9	18.5
GET-T [20]	47.9	16.8	48.4	18.2	21.9
S5-ViT-B [34]	47.4	8.2	47.2	9.6	17.5
ASTMNet [16]	46.7	35.6	48.3	72.3	> 100
SAST-CB [21]	48.2	-	48.7	57.5 (19.7)	18.9
ERGO-12 [33]	50.4	69.9	40.6	100.0	59.6
RT-DETR-T (ours)	46.0	6.4	44.1	8.6	20.1
RT-DETR-B (ours)	47.6	10.5	45.2	14.9	42.8
EvRT-DETR-T (ours)	52.3	8.4	49.9	12.5	34.4
EvRT-DETR-B (ours)	**52.7**	12.7	**50.1**	18.8	57.1

Table 1. **Performance Comparison on Event-based Camera Datasets.** Our approach achieves state-of-the-art results on both Gen1 and 1Mpx benchmarks with significant improvements over previous methods. The first stage of the framework (RT-DETR-T/B) already achieves competitive performance using only image-based detection, while the proposed I2EvDet-adapted models (EvRT-DETR-T/B) surpass all existing approaches. Runtime measurements are reported in milliseconds on an NVIDIA T4 GPU. Values in parentheses represent original reported times on different hardware with T4-equivalent times calculated using FLOPS/throughput conversion.

Evaluation Protocol. We evaluate EvRT-DETR using the standard COCO (Common Objects in Context) Mean Average Precision (mAP) measure [17] that allows direct comparison with existing methods. For consistency with previous works, we use the EBC-specific implementation of the COCO metrics provided by Prophesee's Automotive Dataset Toolbox [22, 23]. All inference times reported in Table 2 are standardized to NVIDIA T4 GPU performance for fair comparison. Inference times include model forward pass and post-processing but exclude event preprocessing, following standard reporting practices. When prior work reported only preprocessing-inclusive times, we use their reported values directly.

4.2. Performance Analysis

Table 1 presents a comprehensive comparison of our EvRT-DETR models against state-of-the-art approaches on the Gen1 and 1Mpx benchmarks.

First, our baseline adaptation using two RT-DETR variants, RT-DETR-T (Tiny, with ResNet-18 backbone) and RT-DETR-B (Base, with ResNet-50 backbone), and trained on simple event frames achieves competitive performance (46.0% and 47.6% mAP on Gen1; 44.1% and 45.2% mAP on 1Mpx) without requiring specialized architectures or complex event representations. This demonstrates that modern mainstream object detectors can be effectively adapted to event data through appropriate training.

More importantly, our complete EvRT-DETR models with temporal adaptation significantly outperform all existing methods, achieving new state-of-the-art results on both benchmarks. EvRT-DETR-B reaches 52.7% mAP on Gen1 (+2.3% over the previous best) and 50.1% mAP on 1Mpx (+1.4% over the previous best). Even our lighter EvRT-DETR-T model surpasses all existing approaches while maintaining faster inference times.

Figure 4 shows how the temporal adaptation module provides crucial capabilities for handling stationary objects – a fundamental challenge for event cameras. When objects are in motion, both RT-DETR and EvRT-DETR perform well, yet, when motion stops and event generation becomes minimal, EvRT-DETRs temporal memory enables consistent detection.

These performance improvements come with minimal computational overhead. While we focus primarily on detection accuracy in this work, the runtime measurements in Table 1 demonstrate that our approach maintains competitive inference speeds despite adding temporal processing capabilities.

4.3. Framework Generalizability

To demonstrate that our results are not specific to the RT-DETR architecture, we apply the I2EvDet framework to YOLOX [7]. YOLOX is a CNN-based detector that contrasts with RT-DETR's transformer architecture and has been used in other EBC applications [9, 34]. Table 2 shows that our temporal adaptation approach consistently

Detector	mAP (%)		
	Original	I2EvDet	Improvement
RT-DETR-T	46.0	52.3	+6.3
RT-DETR-B	47.6	52.7	+5.1
YOLOX-T	36.0	42.4	+6.4
YOLOX-X	43.4	47.8	+4.4

Table 2. **Generalizability of I2EvDet Framework Across Different Object Detection Architectures on the Gen1 Dataset**. Our temporal adaptation consistently improves performance regardless of the base detector architecture.

improves performance across different baseline architectures on the Gen1 dataset. The substantial improvements (4.4-6.4 mAP) across both RT-DETR and YOLOX variants validate that I2EvDet represents a general framework for adapting image-based detectors to temporal domains rather than an architecture-specific optimization. Additional details are provided in the supplementary material.

4.4. Analysis of Adaptation Design Choices

We systematically analyze key design choices in our temporal adaptation module to better understand the factors contributing to EvRT-DETR's performance. This analysis provides insights into efficient adaptation strategies for event-based object detection. Additional ablation studies are provided in the supplementary material.

Model	mAP (%)	mAP_{50} (%)	mAP_{75} (%)
RT-DETR-B (ours)	47.6	76.3	49.5
EvRT-DETR-B $(0, 1, 1)$	51.0	80.7	54.0
EvRT-DETR-B $(1, 0, 1)$	52.2	81.3	55.1
EvRT-DETR-B $(1, 1, 0)$	<u>52.4</u>	<u>81.7</u>	<u>55.3</u>
EvRT-DETR-B $(1, 1, 1)$	**52.7**	**82.0**	**56.0**

Table 3. **Effect of Temporal Module Placement.** Performance on the Gen1 dataset with ConvLSTM modules at different feature scales. The array notation (x, y, z) indicates presence (1) or absence (0) of temporal modules at scales $\mathcal{E}_3$, $\mathcal{E}_4$, and $\mathcal{E}_5$, respectively. Low-level features benefit most from temporal adaptation.

4.4.1. Multiscale Temporal Adaptation

A crucial aspect of our approach is the insertion of temporal modules at multiple scales in the feature hierarchy. Table 3 shows performance when selectively removing ConvLSTM cells from specific feature levels with the array notation (x, y, z) indicating presence (1) or absence (0) at each scale $\{\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5\}$.

The results reveal a clear pattern: the lowest-level features ($\mathcal{E}_3$) benefit most significantly from temporal adaptation as performance drops substantially when this module is removed. The impact diminishes at higher feature levels, suggesting that temporal information is most valuable for fine-grained, lower-level features that capture detailed motion patterns. This insight can guide more efficient adaptation designs, wherein computational resources are directed to where they provide maximum benefit.

Model	mAP (%)	mAP_{50} (%)	mAP_{75} (%)	N_{T} (M)
EvRT-DETR-B (M=64)	52.1	81.3	54.9	2.3
EvRT-DETR-B (M=128)	52.5	81.7	<u>55.6</u>	5.4
EvRT-DETR-B (M=256)	<u>52.7</u>	**82.0**	**56.0**	14.4
EvRT-DETR-B (M=512)	**52.9**	<u>81.9</u>	**56.0**	42.9

Table 4. **Impact of Temporal Module Capacity.** Performance on Gen1 dataset with varying ConvLSTM hidden dimensions (M). The rightmost column shows the number of trainable parameters (in millions) in the temporal adaptation module.

4.4.2. Parameter Efficiency in Adaptation

Inspired by parameter-efficient techniques such as LoRA (Low-Rank Adaptation) [15], we investigate the minimal temporal module capacity needed for effective adaptation. Table 4 shows EvRT-DETR-B performance with varying hidden dimensions in the ConvLSTM modules.

Remarkably, reducing hidden dimensions to 64 features (a 4x reduction from our baseline) results in only modest performance degradation (-0.6% mAP) while requiring just 2.3 M additional parameters – merely 5.4% of the base RT-DETR-B model's 42.8 M parameters. This configuration closely resembles LoRA-style adaptation, achieving substantial performance gains with minimal parameter overhead. Even our most parameter-efficient adaptation (M=64) outperforms all previous state-of-the-art approaches, demonstrating that effective temporal adaptation does not require extensive architectural modifications. Increasing the hidden dimension to 512 features (42.9 M parameters, nearly doubling the model size) yields only a marginal 0.2% mAP improvement, indicating diminishing returns.

5. Conclusions

This work introduces I2EvDet, a framework that efficiently transforms image-based detectors into video-capable models using minimal modifications. Applied to event-based cameras, our approach demonstrates that mainstream object detection architectures can be effectively adapted to temporal domains without specialized redesign. RT-DETR trained on simple event representations achieves performance comparable to specialized methods, while our I2EvDet-adapted model EvRT-DETR achieves state-of-the-art results on Gen1 ($+2.3$ mAP) and 1Mpx ($+1.4$ mAP) benchmarks. Ablation studies provide insights about optimal temporal module design and parameter efficiency. These results suggest that bridging mainstream computer vision and specialized domains can be achieved through targeted adaptation rather than complete architectural redesign, potentially accelerating progress across diverse temporal visual domains.

Acknowledgments

The authors extend their sincere gratitude to Yuewei Lin (Brookhaven Lab) for his valuable feedback and suggestions. This work presented in this paper was funded by the U.S. Department of Energy (DOE), National Nuclear Security Administration, Office of Defense Nuclear Nonproliferation Research and Development. The manuscript has been authored by Brookhaven National Laboratory managed by Brookhaven Science Associates, LLC under the Contract No. DE-SC0012704 with the U.S. Department of Energy.

References

[1] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Gary Cottrell, and Julian J. McAuley. ReZero is all you need: fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, pages 1352–1361. AUAI Press, 2021. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 2, 3

[3] Loïc Cordone, Benoît Miramond, and Philippe Thiérion. Object detection with spiking neural networks on automotive event data. In *International Joint Conference on Neural Networks, IJCNN 2022, Padua, Italy, July 18-23, 2022*, pages 1–8. IEEE, 2022. 3, 7

[4] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *CoRR*, abs/2001.08499, 2020. 2, 6

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 4

[6] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1, 3, 4

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *CoRR*, abs/2107.08430, 2021. 7

[8] Daniel Gehrig and Davide Scaramuzza. Pushing the limits of asynchronous graph-based object detection with event cameras. *CoRR*, abs/2211.12324, 2022. 3

[9] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13884–13893. IEEE, 2023. 3, 4, 6, 7

[10] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014. 3

[11] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Trans. Mach. Learn. Res.*, 2024, 2024. 4

[12] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 4, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2790–2799. PMLR, 2019. 2, 4, 5

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2, 4, 5, 8

[16] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Trans. Image Process.*, 31:2975–2987, 2022. 3, 7

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 7

[18] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII*, pages 415–431. Springer, 2020. 7

[19] Yansong Peng, Yueyi Zhang, Peilin Xiao, Xiaoyan Sun, and Feng Wu. Better and faster: Adaptive event conversion for event-based object detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington,*

DC, USA, February 7-14, 2023, pages 2056–2064. AAAI Press, 2023. 3, 7

[20] Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. GET: group event transformer for event-based vision. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6015–6025. IEEE, 2023. 2, 3, 7

[21] Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16794–16804. IEEE, 2024. 3, 7

[22] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3, 6, 7

[23] Prophesee. Prophesee's automotive dataset toolbox. `https://github.com/prophesee-ai/prophesee-automotive-dataset-toolbox`, 2024. 7

[24] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016. 2, 3

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 2

[26] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. AEGNN: asynchronous event-based graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12361–12371. IEEE, 2022. 3, 7

[27] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810, 2015. 6

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 3

[29] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[30] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on Real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16965–16974. IEEE, 2024. 2, 4, 5

[31] Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 4

[32] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[33] Nikola Zubic, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12800–12810. IEEE, 2023. 2, 3, 7

[34] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 5819–5828. IEEE, 2024. 3, 7