M^{2}Former: Enhancing Event-Based RT-DETR for Robust and Lightweight Space Object Detection

Ruitao Pan, Chenxi Wang, Bin Han, *Student Member, IEEE*, Xinyu Zhang, Zhi Zhai, Jinxin Liu, Naijin Liu, and Xuefeng Chen, *Senior Member, IEEE*

***Abstract*—With increasing human space activities, detecting resident space objects (RSOs) has become critical for space monitoring and on-orbit missions. Traditional optical sensors struggle in space environments due to extreme illumination variations and motion blur. Event cameras, bioinspired sensors that asynchronously record per-pixel brightness changes, offer high temporal resolution, wide dynamic range, and low power consumption, making them promising for orbital sensing yet underexplored in this context. In this work, we present the first systematic study of spaceborne event-based space object detection. To address the scarcity of event data, we construct a large-scale dataset named Event-based SPAcecraft Recognition leveraging Knowledge of Space Environment (E-SPARK) by applying affine transformations and advanced event simulators to existing datasets. Building on this dataset, we propose a lightweight multiscale MetaFormer backbone called M^{2}Former together with an area-aware loss (AAL) tailored for small-object detection. These components are integrated into the real-time detection Transformer (RT-DETR) framework, a Transformer-based detector known for its robustness but higher computational cost compared to you only look once (YOLO) models. Our design reduces parameters and complexity by over 50% while maintaining comparable detection accuracy. In addition, we design an improved data augmentation strategy that enriches supervision density and data diversity, further boosting detection performance. Experiments on both synthetic and real event data demonstrate that our method achieves state-of-the-art performance and strong generalization. These results highlight the potential of event cameras as a reliable sensing modality for spaceborne detection. The dataset, code, and supplementary materials are publicly available at https://iamie-vision.github.io/M2Former/**

***Index Terms*—Event-based vision, multiscale MetaFormer design, real-time detection Transformer (RT-DETR), space object detection.**

I. Introduction

THE rapid growth of human space activities has significantly increased the number of resident space objects (RSOs), escalating collision risks and threatening space operations. Reliable detection of RSOs is essential for a wide range of space missions, such as remote monitoring [1], on-orbit servicing, and active debris removal [2].

Traditional spaceborne detection systems rely on electro-optical sensors such as visible-light cameras, infrared (IR) sensors, and LiDAR. Each of these modalities has contributed significantly to space monitoring and continues to play an important role. Visible-light cameras provide high spatial resolution but can be affected by illumination variability [3]. IR sensors are effective in darkness, but their resolution is typically limited, and they can be influenced by thermal conditions [4]. LiDAR offers precise range measurements but consumes more power and is sensitive to surface reflectivity [5]. Multisensor platforms such as NASA's Raven [6] exemplify how these sensors can be combined to complement one another in practice. Nevertheless, extreme illumination variations and rapid relative motion remain challenging scenarios for conventional systems, motivating the exploration of novel sensing paradigms.

Event cameras, which measure pixel-wise brightness changes asynchronously, represent one such paradigm. By generating sparse event streams with high temporal resolution, wide dynamic range, low latency, and low power consumption [7], they offer capabilities that are naturally suited to demanding space sensing scenarios. Their robustness under extreme lighting and rapid motion has already supported diverse applications, including satellite pose estimation [8], [9], structure-from-motion [10], and star tracking [11]. Several prior studies have also explored event cameras for RSO detection and tracking [12], [13]. However, these works have primarily focused on ground-based telescopes or handcrafted feature-based methods, which face limitations such as low signal-to-noise ratios (SNR) and noisy event streams, thereby restricting their direct applicability to on-orbit platforms. Consequently, event cameras remain underexplored for spaceborne detection, where they could serve as a valuable complement to established electro-optical sensors.

Over the past decade, deep learning techniques have substantially improved object detection performance. You only look once (YOLO) [14] has been widely applied to real-time detection tasks due to its compact architecture and fast inference speed. Detection Transformer (DETR) [15] adopts the Transformer architecture to directly learn object–query relationships and build contextual information in an end-to-end manner, thereby eliminating the need for postprocessing of nonmaximum suppression (NMS), a common heuristic used to filter duplicate bounding boxes in traditional detectors.

Received 21 July 2025; revised 19 October 2025 and 11 November 2025; accepted 19 November 2025. Date of publication 21 November 2025; date of current version 4 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U22B2013. *(Corresponding author: Chenxi Wang.)*

Ruitao Pan, Bin Han, and Xinyu Zhang are with the School of Future Technology, Xi'an Jiaotong University, Xi'an 710049, China, and also with the National Key Laboratory of Aerospace Power System and Plasma Technology, Xi'an Jiaotong University, Xi'an 710049, China.

Chenxi Wang, Zhi Zhai, Jinxin Liu, and Xuefeng Chen are with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the National Key Laboratory of Aerospace Power System and Plasma Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wangchenxi@xjtu.edu.cn).

Naijin Liu is with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with China Academy of Space Technology, Beijing 100094, China.

Digital Object Identifier 10.1109/TGRS.2025.3636122

This makes DETR inherently robust to thresholds sensitive to illumination and motion. Recently, real-time DETR (RT-DETR) [16] has been proposed for real-time detection by combining early stage convolutions with a final Transformer layer to reduce computational complexity. However, its minimal configuration, RT-DETR-R18, remains heavier than lightweight YOLO models, limiting its applicability on resource-constrained devices. Nevertheless, its end-to-end design offers a simple yet robust architecture suitable for challenging space environments.

Deep learning, particularly Transformer-based methods, requires large-scale datasets for effective training. The lack of high-quality, domain-specific datasets for spaceborne detection remains a critical bottleneck. Acquiring real orbital data is challenging, and constructing ground-based testbeds is resource-intensive [17]. Generating data from 3-D game engines offers a viable alternative [18], [19], and existing public datasets primarily consist of synthetic images [20], [21], [22]. However, there is still a lack of large-scale event datasets, which significantly hinders the development of event-based space object detection.

To address the above challenges, we first create a publicly available event dataset called Event-based SPAcecraft Recognition leveraging Knowledge of Space Environment (E-SPARK), generated through affine transformations and advanced event simulators. Then, we propose a lightweight multiscale MetaFormer backbone called M^{2}Former for resource-constrained platforms, where MetaFormer is a Transformer-like architecture that generalizes the design of vision backbones by decoupling token mixing and channel mixing. M^{2}Former integrates efficient convolutional and downsampling operations with multiscale feature extraction capability, the ability to capture objects of different sizes by jointly modeling fine- and coarse-grained representations. In addition, we introduce an area-aware loss (AAL) for small-object detection, which adaptively emphasizes localization errors based on object scale. These components are integrated into the RT-DETR framework to achieve an optimal balance between detection accuracy and computational efficiency. Moreover, we design an improved data augmentation strategy that enriches supervision density and data diversity, thereby facilitating more effective training and improved detection performance. The proposed method not only achieves superior performance on synthetic event data but also demonstrates strong generalization to real event data collected from a ground-based testbed.

As shown in Fig. 1, compared to RT-DETR-R18 and its variants (PoolFormer-S12, RepViT-M0.9, and MobileNetV2), our method achieves a more favorable tradeoff between detection accuracy and model complexity. Furthermore, when trained with the improved data augmentation strategy (denoted as M^{2}Former*), our method surpasses representative YOLO models and outperforms the variant of RT-DETR-R18 trained with the default data augmentation strategy (denoted as RT-DETR-R18*). Our main contributions are summarized as follows.

1) We conduct the first systematic investigation of spaceborne event-based space object detection. Extensive experimental results demonstrate the unique advantages of event cameras for robust spaceborne detection.
2) We create and release an event dataset called E-SPARK for space object detection using simple but effective event simulation techniques.
3) We propose a lightweight backbone called M^{2}Former for resource-constrained platforms and introduce an AAL for small-object detection. These components are integrated into the RT-DETR framework to achieve an optimal balance between detection accuracy and computational efficiency.
4) We design an improved data augmentation strategy that enriches supervision density and data diversity, thereby facilitating more effective training and improved detection performance.
5) The proposed method is evaluated on real event data collected from a ground-based testbed, demonstrating the strong generalization for applications.

The rest of this article is organized as follows. Section II reviews related work on object detection and event-based vision. Section III details the proposed method, including the event data generation pipeline, the model architecture, the loss function, and the data augmentation strategy. Section IV presents experimental results and analysis on synthetic and real event data. Finally, Section V concludes this article and outlines future directions.

II. Related Work

A. Real-Time Object Detectors

Mainstream real-time object detectors include one-stage models such as YOLO [14], single shot multibox detector (SSD) [23], and EfficientDet [24], which are widely adopted in resource-constrained scenarios due to their compact architectures and high inference speed. The YOLO family has evolved from anchor-based designs, such as YOLOv5 [25], which rely on predefined bounding boxes (anchors) to guide detection, to more advanced anchor-free designs like YOLOv8 [26] and YOLOv11 [27], which directly predict object locations without using anchor priors. Such models have been explored in various space applications. For example, EfficientDet is applied to satellite localization and classification tasks for orbital situational awareness [28]. YOLOv5 with attention mechanisms is utilized for multiscale satellite detection [29]. Semi-supervised detection methods are developed for identifying close-range spacecraft and debris with limited annotations [30]. However, most of these approaches rely on NMS [31], a heuristic postprocessing step that removes duplicate detections but is sensitive to threshold tuning and can affect stability in harsh space environments.

To address this limitation, DETR [32] introduces a Transformer-based architecture that performs one-to-one NMS-free assignment via the Hungarian algorithm [33] for bipartite matching of predictions and ground truth. Successive improvements such as Deformable DETR [34], dynamic anchor boxes DETR (DAB-DETR) [35], denoising DETR (DN-DETR) [36], and DETR with improved denoising anchor boxes (DINO) [37] progressively enhance multiscale perception, object query modeling, and training convergence.

RT-DETR builds upon these developments by employing an efficient encoder that primarily uses convolutional layers for early stage feature extraction, while integrating attention mechanisms at the final stage to retain global modeling capability [16]. Although this hybrid design improves efficiency compared to earlier DETR variants, even the smallest RT-DETR configuration still incurs higher parameter counts and computational overhead than lightweight YOLO models, thereby limiting its deployment on resource-constrained platforms.

Recent studies [38], [39] suggest that the success of Transformer-based architectures largely derives from the MetaFormer macrostructure, which consists of token mixers and channel mixers. Inspired by RepViT [40], which offers an efficient convolutional reinterpretation of this paradigm, we introduce M²Former, a lightweight MetaFormer-style backbone that enables efficient spatial and channel-wise feature modeling while maintaining the efficiency of convolutional networks. Our redesigned RT-DETR backbone preserves detection accuracy while significantly reducing model parameters and computational complexity, making it highly suitable for spaceborne deployment.

B. Event-Based Object Detectors

Event cameras offer inherent advantages under challenging conditions where conventional frame-based cameras are susceptible to extreme illumination and motion blur. These conditions are commonly encountered in space environments, making event cameras a promising sensing modality for space object detection. Early studies have shown that ground-based event cameras can support RSO detection and tracking, with tailored datasets and feature-based algorithms [12], [13]. While these works demonstrate the potential of event cameras, challenges such as low SNR and noisy event streams limit their direct applicability to on-orbit platforms, motivating the need for more robust and scalable detection methods.

Beyond space applications, event-based object detection methods can be broadly categorized into two paradigms. The first directly processes sparse event streams using spiking neural networks (SNNs) [41], [42], [43] or graph neural networks (GNNs) [44], [45], [46], which preserve temporal precision and energy efficiency but often suffer from unstable training and limited compatibility with mainstream deep learning toolchains. The second paradigm converts asynchronous events into dense representations, such as event histograms [47], time surfaces [48], or event volumes [49], which are then processed by 2-D/3-D convolutional neural networks (CNNs) [50], [51], Transformers [52], [53], or recurrent neural networks (RNNs) [54], [55]. These approaches integrate more seamlessly with existing architectures, though they typically sacrifice temporal resolution due to fixed aggregation windows.

In this work, we focus on dense event representations. Compared with existing datasets, space imaging covers vast scenes in which RSOs often appear as small objects with sparse event activity. Moreover, current detection frameworks generally lack effective multiscale feature extraction. These limitations reduce their applicability to spaceborne detection. To address these challenges, we propose M²Former, a backbone with enhanced multiscale capability, and introduce an AAL to improve the localization of small targets, thereby advancing event-based detection in orbital environments.

C. Event Datasets and Simulators

The training of event-based object detectors heavily relies on high-quality annotated datasets. Public event datasets such as N-Caltech101 [56], MVSEC [57], and Gen1 [58] primarily focus on tasks such as classification, simultaneous localization and mapping (SLAM), and autonomous driving. There is still a lack of domain-specific event datasets for spaceborne detection.

To address the scarcity of event data, simulators such as an open event camera simulator (ESIM) [59] and from video frames to realistic DVS events (v2e) [60] are commonly used to convert videos into synthetic event streams by monitoring pixel intensity changes induced by motion, with noise models added to improve realism. However, these simulators require video inputs and are not applicable to static images. Some generative learning methods [61], [62], [63] have been proposed to generate events from images via domain adaptation techniques [64], [65]. Nevertheless, these approaches typically depend on large-scale training data from both source and target domains, which is difficult to obtain in the context of space applications. As an alternative, Cao et al. [66] proposed a method for generating events from static images using artificial optical flow. Although this approach eliminates the need for video input, it often produces unrealistic motion patterns due to the random direction and magnitude of the synthetic optical flow.

To overcome these limitations, we propose a flexible data synthesis method that applies affine transformations to static images to generate image sequences. Then, physically consistent event streams are produced using advanced event simulators with the image sequences as input. This approach introduces controllable motion while maintaining realistic event characteristics, enabling the creation of large-scale event datasets for space object detection.

III. Method

A. Event Data Generation

1) Image-to-Event Transfer: To address the shortage of event data for space object detection, we generate synthetic event streams from static images using motion-based simulation. Given a static image $I(x, y)$, we construct an image sequence $\{I_t\}_{t=1}^{T}$ by applying affine transformations, including translation, scaling, and rotation. The incremental motion parameters are empirically selected to ensure soft event generation: translation offsets are sampled from $[0.5, 0.8]$ pixels per step, the scaling factor from $[1.001, 1.003]$, and the rotation angle from $[0.05, 0.08]$. These transformations accumulate across $N = 6$ frames to approximate a smooth virtual camera trajectory. The parameters are chosen based on whether the distribution and intensity of events in the generated frames appear reasonable, avoiding both excessive transformations that cause event trailing noise and overly weak transformations that lead to sparse or insufficient events. The transformation matrix $\mathbf{A}_t \in \mathbb{R}^{2\times 3}$ at each step is defined as

$$\mathbf{A}_t = \mathbf{R}(\theta_t) \cdot \mathbf{S}(s_t) + \mathbf{T}(\Delta x_t, \Delta y_t) \quad (1)$$

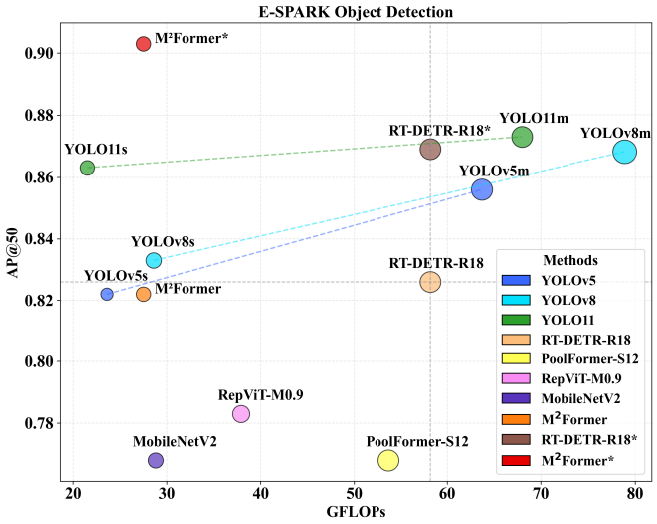

Fig. 1. Performance comparison on the E-SPARK dataset. AP@50 refers to the mean average precision (AP) at an IoU threshold of 0.5, and GFLOPs denotes the giga floating-point operations used to measure computational complexity. The size of each circle indicates the number of model parameters. M^{2}Former* denotes the variant of M^{2}Former trained with the improved data augmentation strategy, and RT-DETR-R18* denotes the variant of RT-DETR-R18 trained with the default data augmentation strategy.

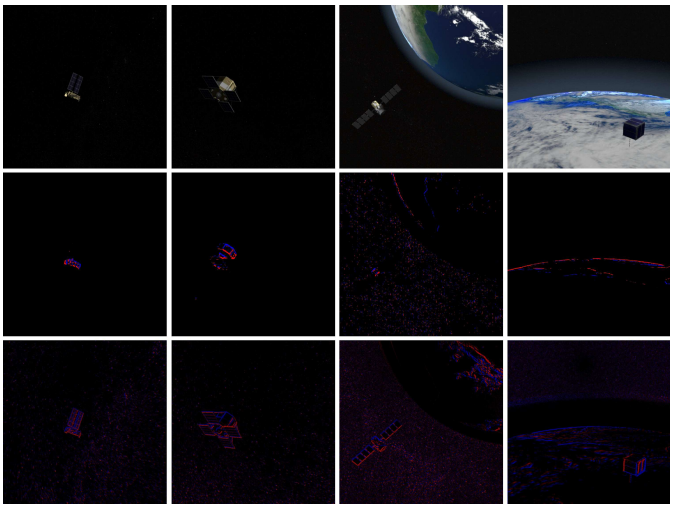

Fig. 2. Comparison of event generation methods: (top) RGB inputs, (middle) Cao et al.'s method [66], and (bottom) our method. Events are visualized as RGB images, where positive events are drawn in red and negative events are drawn in blue.

where θ_t is the rotation angle; s_t is the scaling factor; Δx_t and Δy_t are translation offsets; and $\mathbf{R}, \mathbf{S}$, and $\mathbf{T}$ denote rotation, scaling, and translation matrices, respectively.

The resulting image sequences are then fed into the v2e simulator [62], which generates events by computing logarithmic intensity differences between consecutive frames. The simulator is configured with contrast thresholds (± 0.15) to control event triggering sensitivity, Gaussian smoothing ($\sigma = 0.03$) to suppress pixel-level noise, a low-pass cutoff frequency of 30 Hz to filter high-frequency artifacts, and an exposure duration of 5 ms per frame at 100-Hz input frame rate, under a spatial resolution of 640 $\times$ 480. The output events are subsequently converted into dense representations for training.

This pipeline allows us to leverage image datasets to generate high-quality event datasets. Unlike the method of Cao et al. [66], which synthesizes artificial optical flow with randomly sampled directions and magnitudes, our approach is deterministic and controllable, ensuring physically consistent event data. As shown in Fig. 2, our method has two key advantages. First, in Cao et al.'s approach, the optical flow intensity is randomly sampled. When the sampled intensity is weak, the resulting event activity becomes negligible, leading to information loss at object boundaries, where only partial events are triggered. In contrast, our method controls the event density through the magnitude of the affine transformations, producing denser and more informative events. Second, because the optical flow direction in Cao et al.'s method is randomly assigned, the generated motion patterns are often inconsistent. Our method, on the other hand, produces smoother and more realistic camera trajectories, thereby preserving physically consistent motion cues.

We applied this pipeline to the SPARK2021 dataset [21], which consists of 150 000 RGB images across 11 object categories. Using the proposed simulation method, we generated the corresponding event dataset called E-SPARK. Further details are provided in Section IV-A.

2) Event Representation: Event cameras operate fundamentally differently from conventional frame-based cameras. Instead of capturing images at fixed intervals, they asynchronously record changes in logarithmic brightness at each pixel. An event at pixel location (x, y) and timestamp t is triggered when the change in log intensity ΔL exceeds a predefined threshold θ, which is formulated as

$$\Delta L = \log I(x, y, t) - \log I(x, y, t - \delta t) \geq p \cdot \theta \tag{2}$$

where $I(x, y, t)$ denotes pixel intensity at time t, $p \in \{+1, -1\}$ represents the event polarity (indicating a brightness increase or decrease), and θ is the contrast sensitivity threshold.

Each event is represented as a tuple $e_i = (x_i, y_i, t_i, p_i)$. Raw event streams are inherently sparse and asynchronous, making them incompatible with standard deep neural networks that require dense, grid-based inputs. To enable effective training, events within a fixed time window $[t, t + \Delta t]$ are aggregated into dense representations, including event histogram [see (3)], time surface [see (4)], and event volume [see (5)], defined as follows:

$$H(x, y) = \sum_{i=1}^{N} p_i \cdot \delta(x_i, y_i) \tag{3}$$

where events are accumulated at each pixel according to their polarity p_i and $\delta(\cdot)$ denotes the Kronecker delta function

$$T(x, y) = p_i \cdot \exp\left(-\frac{|t_0 - t(x_i, y_i)|}{\tau}\right) \tag{4}$$

where $t(x, y)$ denotes the timestamp of the latest event at pixel (x, y), t_0 is the reference time (e.g., start of the window), and τ controls the exponential decay rate

$$V(x, y, p, k) = \sum_{i=1}^{N} \delta(x_i, y_i, p_i) \cdot 1_{[t_k, t_{k+1})}(t_i). \tag{5}$$

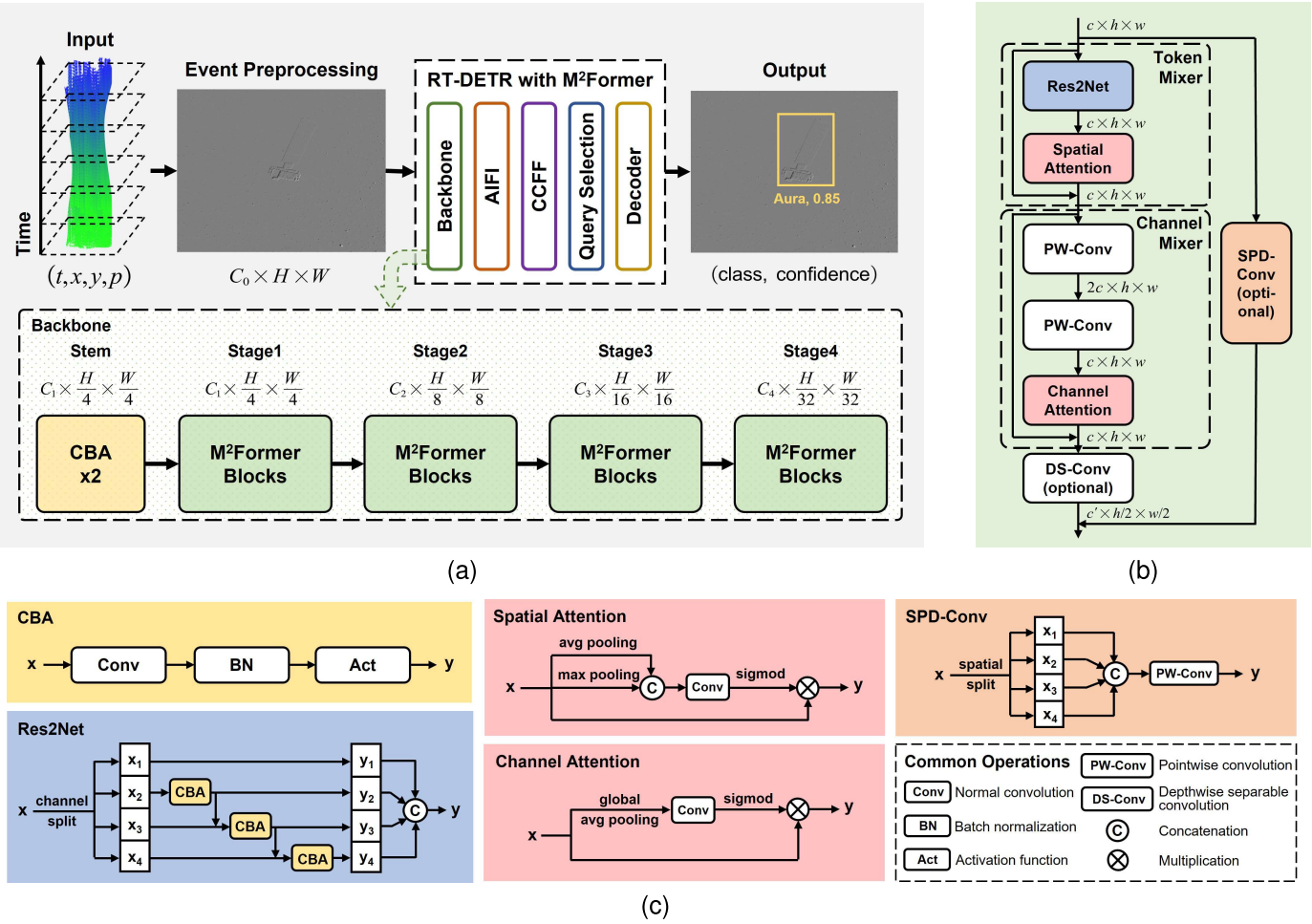

Fig. 3. Architecture of the proposed method. (a) Overall framework. (b) M²Former block. (c) Module design.

Here, the time window is evenly divided into K temporal bins $[t_k, t_{k+1})$. The indicator function $1_{[t_k,t_{k+1})}(t_i)$ equals 1 if the event timestamp t_i falls within the kth bin and 0 otherwise. This representation finally constructs a 4-D tensor $V \in \mathbb{R}^{H\times W\times 2\times K}$.

To facilitate stable training, all three representations are normalized during data preprocessing. Specifically, event histograms are scaled separately per polarity to the range $[-1, 1]$; time surfaces inherently fall within $[-1, 1]$ due to the exponential function; and event volumes are normalized per polarity channel to the range $[0, 1]$. These steps help mitigate intensity disparities across different samples and improve model convergence.

B. Model Architecture

The encoder of RT-DETR consists of a CNN-based ResNet backbone for early stage feature extraction, followed by an attention-based intrascale feature interaction (AIFI) module that enhances contextual modeling. In addition, a CNN-based cross-scale feature fusion (CCFF) module is employed to aggregate information across different resolutions, further improving the encoder's multiscale feature extraction capabilities. Although this hybrid design reduces the computational complexity compared to a fully Transformer-based encoder, the minimal configuration of RT-DETR (i.e., RT-DETR-R18) remains heavier than lightweight YOLO models, limiting its applicability on resource-constrained platforms.

Since the original RT-DETR backbone costs considerable computation, our improvements focus on the redesign of the backbone. We propose M²Former, a lightweight CNN backbone inspired by the MetaFormer paradigm and optimized for real-time applications. As shown in Fig. 3(a), the overall framework integrates the redesigned backbone with the original RT-DETR components. While maintaining the structure of ResNet, M²Former replaces standard residual blocks with novel M²Former blocks. Initially, a stem consisting of two CBA layers is applied to extract features with spatial downsampling at a stride of 4. This design aligns with the characteristics of event data, which contains texture information analogous to the low-level features captured by early convolutional layers. As a result, critical information is preserved despite resolution reduction. Subsequently, a sequence of M²Former blocks extracts features at increasing strides (8, 16, and 32), producing robust multiscale representations. Similar to the original RT-DETR, the final feature map is fed into the AIFI module, where the attention mechanism is applied to the lowest resolution features. Multiscale features are then fused through the CCFF module. After query selection, a Transformer-based decoder is employed to produce detection

outputs in an NMS-free manner. Specifically, as illustrated in Fig. 3(b), M^{2}Former block consists of a token mixer, a channel mixer, residual connections, and an optional SPD-Conv downsampling path, with detailed shown in Fig. 3(c).

The token mixer captures spatial information with a Res2Net module [67], followed by a spatial attention module. For the Res2Net module, the input feature map is divided into multiple channel groups, which are hierarchically processed via residual connections. Each group integrates contextual information from preceding groups through a 3×3 convolution. This structure enables progressive encoding of both fine-grained textures and broader spatial context, resulting in efficient multiscale feature representation. Following hierarchical fusion, the spatial attention module inspired by CBAM [68] is applied. It concatenates average- and max-pooled feature maps along the channel dimension, processes them using a 7×7 convolution and sigmoid activation, and produces spatial attention weights. These weights highlight informative spatial regions, which are particularly beneficial for sparse event data.

The channel mixer, located after the token mixer, consists of a two-layer pointwise convolutional multilayer perception (MLP) and a channel attention mechanism. The MLP first expands and then compresses the channel dimensions, enhancing nonlinear modeling of interchannel relationships. The channel attention module, inspired by efficient channel attention (ECA) [69], captures cross-channel dependencies using a 1-D convolution applied to globally average-pooled descriptors, where the kernel size is adaptively selected based on the number of input channels. The resulting attention weights emphasize informative channels while suppressing redundant responses. Identity residual connections are used in both the token and channel mixers to facilitate gradient flow and stabilize training.

To further reduce computational overhead while retaining detailed features, we adopt depthwise separable convolution for downsampling in the main branch. Meanwhile, the residual branch is implemented using SPD-Conv [70], a spatial-pixel decomposition strategy that splits feature maps along spatial dimensions before channel-wise concatenation and projection. This technique preserves fine details during downsampling and helps mitigate information loss in early layers, which is critical for small-object detection.

In summary, the M^{2}Former backbone adopts a compact architectural design combined with a multiscale feature extraction module to achieve high detection accuracy. Moreover, its efficient attention mechanisms and customized downsampling strategies ensure lightweight computation, making it well-suited for deployment in spaceborne systems with stringent computational constraints.

C. Loss Function

RT-DETR employs varifocal loss (VFL) [71] for classification, which dynamically aligns predicted confidence with localization quality. VFL is formulated as

$$\mathcal{L}_{\text{VFL}}(p,q) = \begin{cases} -q\left(q\log(p) + (1-q)\log(1-p)\right), & q > 0 \\ -\alpha p^{\gamma}\log(1-p), & q = 0 \end{cases} \tag{6}$$

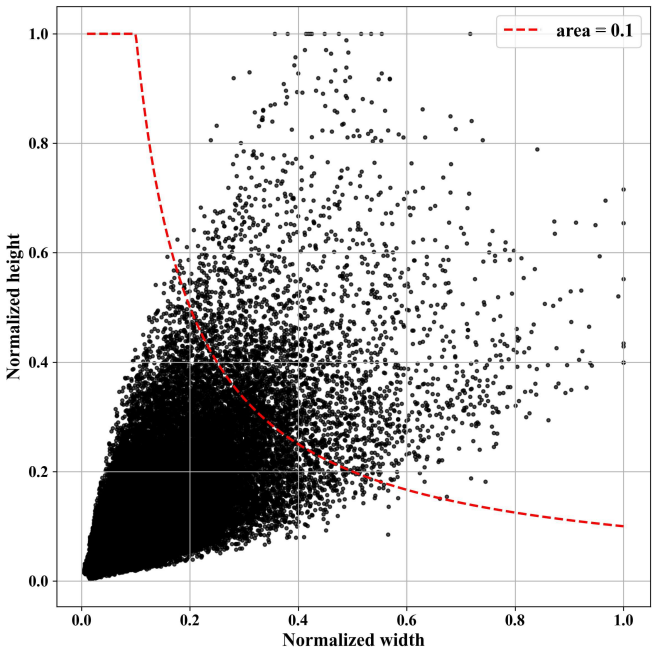

Fig. 4. Normalized object scale distribution in SPARK2021. The red dashed line indicates the area threshold of 0.1 for defining small objects.

TABLE I
DATA AUGMENTATION CONFIGURATIONS

Augmentation	Magnitude	Probability
Mosaic	-	0.5
Mixup	-	0.5
Translation	$\pm$0.1	1.0
Scale	$\pm$0.1	1.0
Horizontal Flip	-	0.5

where p is the predicted confidence, q is the target score defined as the intersection over union (IoU) between the predicted bounding box and the ground truth for the foreground class (and 0 otherwise), α is a balancing factor controlling the weight of negative samples (default $\alpha = 0.75$), and γ is a focusing parameter (default $\gamma = 2.0$). The IoU measures the overlap between two bounding boxes and is computed as the ratio of their intersection area to their union area.

Along with VFL for classification, RT-DETR employs $\mathcal{L}_1$ loss and the generalized IoU (GIoU) loss [72] for localization. Given a predicted bounding box $B_p = (x_p, y_p, w_p, h_p)$ and a ground-truth box $B_t = (x_t, y_t, w_t, h_t)$, $\mathcal{L}_1$ loss and GIoU are defined as

$$\mathcal{L}_1(B_p, B_t) = |x_p - x_t| + |y_p - y_t| + |w_p - w_t| + |h_p - h_t| \tag{7}$$

$$\text{GIoU}(B_p, B_t) = \frac{|B_p \cap B_t|}{|B_p \cup B_t|} - \frac{|C \setminus (B_p \cup B_t)|}{|C|} \tag{8}$$

where $B_p \cap B_t$ and $B_p \cup B_t$ are the intersection and union of B_p and B_t, respectively; C is the smallest enclosing box that completely contains both B_p and B_t; $C \setminus (B_p \cup B_t)$ indicates the region inside the enclosing box C that is not covered by either B_p or B_t; and $|\cdot|$ denotes the area of a region.

Although GIoU can provide gradients even when there is no overlap, its effectiveness heavily depends on the degree of overlap between boxes, making it particularly sensitive to small objects. The normalized Wasserstein distance (NWD) [73] can provide smooth and continuous supervision by jointly modeling discrepancies in the center positions and sizes of bounding boxes. However, the standard NWD formulation treats all objects equally, regardless of their scale, and thus fails to offer the precision required for accurate localization of small objects.

To overcome this issue, we introduce an area-aware reweighting strategy that adaptively adjusts the influence of NWD based on object scale, thereby enhancing localization accuracy for small objects. We first compute the squared Wasserstein-2 distance

$$\mathcal{W}_2^2\left(b_p, b_t\right) = \left(x_p - x_t\right)^2 + \left(y_p - y_t\right)^2 + \frac{1}{4}\left[\left(w_p - w_t\right)^2 + \left(h_p - h_t\right)^2\right]. \quad (9)$$

To introduce scale sensitivity, we define an adaptive normalization factor based on the area of the ground-truth bounding box:

$$\alpha(s) = \frac{C}{2\sigma\left(\log \frac{1}{s}\right)} \quad (10)$$

where $s = w_t \cdot h_t$, $\sigma(\cdot)$ denotes the sigmoid function, and C is an empirically constant (set to 12.8).

The resulting adaptive NWD is defined as

$$\mathrm{NWD} = \exp\left(-\frac{\sqrt{\mathcal{W}_2^2}}{\alpha(s)}\right). \quad (11)$$

This formulation increases gradient contributions for small objects by reducing the normalization factor $\alpha(s)$, thereby enabling better localization of fine-grained targets.

Finally, we define the AAL by combining the GIoU loss and the adaptive NWD loss

$$\mathcal{L}_{\mathrm{AAL}} = \beta \cdot (1 - \mathrm{GIoU}) + (1 - \beta) \cdot (1 - \mathrm{NWD}) \quad (12)$$

where $\beta \in [0, 1]$ balances the two terms. During training, AAL is computed for all positive matches and normalized by the number of matched ground-truth boxes, ensuring stable gradient contributions.

The overall training loss is a weighted sum of the classification loss and the localization loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{VAL}} + 5\mathcal{L}_1 + 2\mathcal{L}_{\mathrm{AAL}}. \quad (13)$$

By integrating this scale-sensitive design into the loss function of RT-DETR, we significantly improve the localization of small objects frequently encountered in large-scale space scenarios.

D. Data Augmentation

Data augmentation plays a vital role in improving the generalization and convergence of detection models, particularly in scenarios with sparse supervision, such as event-based space object detection. Motivated by DEIM [74], which highlights that Mosaic augmentation alleviates sparse matching issues in DETR-based architectures, we design a composite augmentation pipeline to enrich both supervision density and data diversity.

Our augmentation strategy comprises a sequence of transformations, including Mosaic, MixUp, and geometric transformations (translation, scaling, and horizontal flip). Among these, Mosaic increases the number of positive samples by combining multiple images, while MixUp enhances regularization by interpolating supervision signals across samples. These augmentations are especially beneficial for dense one-to-one (O2O) matching and small-object detection, where event activity is often sparse. When combined with geometric transformations, the overall strategy not only provides denser supervision but also increases data variation, thereby supporting more effective training.

Table I summarizes the augmentations used in our pipeline. Each transformation is applied independently with a probability, and its magnitude is sampled uniformly within a specified range.

IV. Experiments

A. Dataset Description

SPARK2021 [21] is a large-scale public dataset for space object detection. It consists of 150 000 high-resolution (1024 × 1024) RGB images generated by the Unity game engine. The dataset includes 11 object categories: 10 satellites (AcrimSat, Aquarius, Aura, Calipso, CloudSat, CubeSat, Jason, Sentinel-6, Terra, and TRMM) and one debris class. Each satellite class contributes 12 500 images, while the debris class contains 25 000 images, yielding a balanced distribution. The dataset is randomly split into training, validation, and test sets with a ratio of 3:1:1.

SPARK2021 simulates diverse orbital environments, sensor configurations, and illumination conditions, supporting the development and evaluation of robust space object detection algorithms. It also presents significant challenges, including varying illumination, motion blur, and a high proportion of small objects. As shown in Fig. 4, most instances fall below the normalized area of 0.1, motivating the design of the M²Former and the AAL.

To enable event-based detection, we construct E-SPARK from SPARK2021 using the proposed event generation method. It applies controlled affine transformations to static RGB images to simulate motion, which are then processed by the event simulator to generate physically consistent event streams. E-SPARK is publicly available to facilitate further research in event-based space object detection.

B. Model Training and Evaluation Metrics

We adopt RT-DETR with ResNet-18 (RT-DETR-R18) as the baseline, which is the minimal configuration of RT-DETR variants. The proposed method is built upon this configuration. To align with the structure of ResNet-18, the number of M²Former blocks is set to $N = 2$ in each stage.

All methods are implemented by PyTorch and trained on an NVIDIA RTX 4090 GPU. The training configurations differ slightly between YOLO models and DETR models.

Both types of models are trained using the Ultralytics training pipeline with some adjustments. For YOLO models, we use the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005. The batch size is set to 64, and the models are trained for 200 epochs. For DETR models, we use the AdamW optimizer with an initial learning rate of 0.0001, a momentum of 0.9, and a weight decay of 0.0001. The batch size is set to 16, and the models are trained for 50 epochs. A linear learning rate schedule is applied. Additionally, the weighting factor of $\lambda = 0.3$ is applied to AAL.

Model performance is evaluated using standard COCO metrics. Specifically, the mean AP at an IoU threshold of 0.5 (AP@50) and the mean AP averaged over IoU thresholds from 0.5 to 0.95 (AP@50:95) are used as the evaluation criteria. To further assess the performance of small-object detection, we report the mean AP for small objects (denoted as APs). To comprehensively evaluate deployment feasibility, the computational complexity is measured in giga floating-point operations (GFLOPs), and the number of parameters is reported to quantify the model's size. Additionally, we employ the Python binding for the NVIDIA Management Library (*pynvml*) to monitor GPU power usage and use the Python *time* library to compute latency. These metrics together provide a comprehensive assessment of both detection performance and practical efficiency on resource-constrained platforms.

C. Analysis of Event Representations

To analyze the impact of different input representations, we evaluate three widely used methods: event histogram, time surface, and event volume. They are evaluated across YOLOv5, YOLOv8, YOLO11, and RT-DETR-R18 using the E-SPARK dataset with original event resolution of 640×480 (i.e., DVXplorer camera configuration). Detection results are summarized in Table II.

TABLE II
PERFORMANCE COMPARISON OF EVENT REPRESENTATIONS

Method	Representation	AP@50	AP@50:95
YOLOv5s	Event Histogram	0.822	0.690
	Time Surface	0.819	0.682
	Event Volume	**0.850**	**0.711**
YOLOv5m	Event Histogram	0.856	0.727
	Time Surface	0.848	0.715
	Event Volume	**0.861**	**0.728**
YOLOv8s	Event Histogram	**0.833**	**0.701**
	Time Surface	0.813	0.680
	Event Volume	0.831	0.699
YOLOv8m	Event Histogram	0.868	0.737
	Time Surface	0.843	0.714
	Event Volume	**0.871**	**0.743**
YOLO11s	Event Histogram	**0.863**	**0.726**
	Time Surface	0.839	0.708
	Event Volume	0.859	0.725
YOLO11m	Event Histogram	**0.873**	**0.742**
	Time Surface	0.856	0.725
	Event Volume	0.865	0.739
RT-DETR-R18	Event Histogram	**0.826**	**0.680**
	Time Surface	0.801	0.663
	Event Volume	0.809	0.659

Note: Bold indicates the highest score among representations for each model, and underline indicates the second-highest score.

In terms of detection accuracy, both event histogram and event volume outperform time surface across most models. Specifically, event volume achieves the highest AP@50 on YOLOv5s, YOLOv5m, and YOLOv8m, whereas event histogram performs best on YOLOv8s, YOLO11s, YOLO11m, and RT-DETR-R18. The performance gap between histogram and volume is consistently small, suggesting that both representations are effective under the current experimental settings.

The superior performance of event histogram and event volume can be attributed to their ability to preserve the spatial structure and polarity information inherent in event data. Event volume explicitly separates positive and negative polarities and distributes them across multiple temporal bins, which in theory enables richer temporal modeling. However, in this study, since the event streams are generated from static images using synthetic motion, the underlying temporal information is limited. As a result, event volume cannot fully leverage its advantages, yielding comparable performance to the event histogram, which accumulates events by polarity without temporal stacking. In contrast, time surface consistently underperforms across all models. This may be due to its reliance on temporal decay. Such decay may suppress useful features, especially for small objects where spatial cues are already sparse.

Among three representations, the event histogram provides the most efficient preprocessing pipeline, as it generates a two-channel tensor without requiring temporal binning. It also achieves competitive detection accuracy across all detectors and integrates seamlessly with standard deep learning frameworks, making it a practical choice for event-based space object detection. Given its favorable tradeoff between accuracy and efficiency, the event histogram is adopted as the default input representation in all subsequent experiments.

D. Analysis of the Proposed Method

To enhance event-based RT-DETR for space object detection, we introduce three major improvements: a lightweight backbone, M^{2}Former, an adaptive localization loss, AAL, and an improved data augmentation strategy. This section presents evaluations of their individual and combined contributions.

We first analyze the superiority of M^{2}Former by replacing the backbone in RT-DETR-R18 with several lightweight alternatives, including PoolFormer-S12 [39], RepViT-M0.9 [40], and MobileNetV2 [75]. PoolFormer and RepViT are both derived from the MetaFormer paradigm; PoolFormer employs pooling operations as its sole token-mixing operator, while RepViT designs a lightweight CNN backbone from a Transformer perspective for mobile deployment. In contrast, MobileNetV2 represents a classical convolutional

TABLE III
DETECTION PERFORMANCE ON EVENT HISTOGRAM INPUTS

Method	AP@50	AP@50:95	APs
YOLOv5s	0.822	0.690	0.458
YOLOv5m	0.856	0.727	0.508
YOLOv8s	0.833	0.701	0.453
YOLOv8m	0.868	0.737	0.512
YOLO11s	0.863	0.726	0.490
YOLO11m	0.873	0.742	0.512
RT-DETR-R18	0.826	0.680	0.499
PoolFormer-S12	0.768	0.610	0.420
RepViT-M0.9	0.783	0.624	0.447
MobileNetV2	0.768	0.623	0.431
M²Former	0.822	0.663	0.466
RT-DETR-R18 + AAL	0.833	0.694	0.516
RT-DETR-R18 + AAL + Aug1	0.869	0.710	0.543
RT-DETR-R18 + AAL + Aug2	0.883	0.729	0.553
M²Former + AAL	0.826	0.677	0.511
M²Former + AAL + Aug1	0.850	0.713	0.559
M²Former + AAL + Aug2	**0.903**	**0.743**	**0.580**

Note: Underline indicates the best score within each framework, and bold highlights the best score across all methods. Aug1 refers to the default data augmentation strategy used in RT-DETR, while Aug2 denotes the improved data augmentation strategy proposed in this work.

TABLE IV
MODEL EFFICIENCY COMPARISON

Method	Params (M)	GFLOPs	Speed (FPS)	Energy (mJ)
YOLOv5s	**9.1**	23.6	**195.1**	21.5
YOLOv5m	20.9	63.7	149.2	67.5
YOLOv8s	11.1	28.6	188.3	**20.3**
YOLOv8m	25.9	78.9	144.9	82.1
YOLO11s	9.4	**21.5**	188.7	28.2
YOLO11m	20.1	68.0	141.9	80.1
RT-DETR-R18	20.1	58.2	98.7	58.1
PoolFormer-S12	20.3	53.7	91.4	89.0
RepViT-M0.9	13.6	37.9	96.4	54.3
MobileNetV2	10.6	28.8	99.5	40.0
M²Former	9.7	27.5	104.3	31.1

Note: Bold denotes the best performance in each column. Inference speed (FPS) and energy consumption (mJ/image) are measured on an NVIDIA RTX 4090 GPU.

network optimized for resource-constrained platforms. As shown in Table III, M²Former achieves 0.822 AP@50, 0.663 AP@50:95, and 0.466 APs, which are notably higher than all other backbones. While PoolFormer-S12, RepViT-M0.9, and MobileNetV2 exhibit significantly lower accuracy, M²Former slightly lags behind RT-DETR-R18. These results demonstrate the strength of M²Former in detection performance. This is attributed to its ability to extract multiscale features and emphasize features through the integration of spatial and channel attention mechanisms. These components jointly highlight salient events while suppressing background noise, which is essential for accurate detection in sparse event inputs.

In terms of efficiency, Table IV shows that M²Former has only 9.7 M parameters and 27.5 GFLOPs, compared to 20.1 M parameters and 58.2 GFLOPs for RT-DETR-R18, representing over a 50% reduction in computational cost. Moreover, M²Former achieves 104.3 FPS with an energy consumption of 31.1 mJ per image, which is faster and more efficient than RT-DETR-R18 (98.7 FPS and 58.1 mJ). Compared to YOLO models, M²Former is lighter than YOLOv8s (11.1 M), close to YOLOv5s (9.1 M), and slightly heavier than YOLO11s (9.4 M). Regarding inference speed, M²Former is slower than the small YOLO variants (YOLOv5s: 195.1 FPS, YOLOv8s: 188.3 FPS, and YOLO11s: 188.7 FPS) but offers competitive runtime and significantly lower energy consumption compared with medium-scale YOLO models (YOLOv8m: 144.9 FPS and 82.1 mJ; YOLO11m: 141.9 FPS and 80.1 mJ). Since real-time detection generally requires only 30 FPS, all tested models already meet this requirement by a wide margin. Further speedups and energy reductions via deployment optimizations are feasible but beyond the scope of this work. Our focus is on lightweight design under the MetaFormer paradigm, a key requirement for resource-constrained spaceborne platforms. In addition, unlike YOLO detectors, our approach is NMS-free, avoiding IoU-threshold sensitivity and yielding more stable behavior under uncertain and dynamic space environments.

Next, we evaluate the impact of the proposed AAL, which adaptively reweights the localization loss based on object scale. This formulation improves gradient flow for small objects, which often appear in space environments. As shown in Table III, incorporating AAL into M²Former improves AP@50:95 from 0.663 to 0.677, and APs from 0.466 to 0.511. The improvements in APs are more obvious, demonstrating the effectiveness of AAL. Similar gains are observed for RT-DETR-R18, confirming the generality of the loss function. These consistent improvements validate the effectiveness of scale-aware supervision in addressing small-object localization, especially under sparse and low-contrast event data.

Then, we assess the contribution of the improved data augmentation strategy, which integrates Mosaic, MixUp, and geometric transformations (denoted as Aug2). It is important to note that RT-DETR-R18 is originally trained with its default augmentation, including multiscale inputs and geometric transformations (denoted as Aug1). As shown in Table III, before applying data augmentations, M²Former performs slightly below RT-DETR-R18. When trained with Aug1 or Aug2, both frameworks exhibit improvements, but Aug2 brings much larger gains. On average, Aug2 improves RT-DETR-R18 by about 2%, while M²Former benefits more substantially with an improvement of about 5%. This indicates that M²Former benefits more from the richer and denser supervision provided by Aug2. With both AAL and Aug2, M²Former achieves the best accuracy, outperforming all other configurations. These results clearly demonstrate the superiority of the proposed augmentation strategy. The combined use of Mosaic and MixUp introduces more diverse and denser supervision and improves the model's ability to detect small and weak events.

Further to demonstrate the superiority of our method, qualitative results are shown in Fig. 5. The results from M²Former most closely align with the ground truth, demonstrating strong classification and localization accuracy. In contrast, RT-DETR-R18 exhibits missed detections, while YOLO11m performs relatively well with fewer false detections. These qualitative

Fig. 5. Detection results of (top) YOLO11m, (middle) RT-DETR-R18, and (bottom) proposed M^{2}Former on the E-SPARK dataset. Events are visualized as grayscale images. Bounding boxes indicate detected targets with class labels and confidence scores. The ground-truth labels for columns 1–4 are Aura, Terra, CloudSat, and Debris, respectively.

TABLE V
ABLATION STUDY ON M^{2}FORMER COMPONENTS

Method	AP@50	AP@50:95
M^{2}Former (baseline)	0.822	0.663
w/o Res2Net	0.795 (-0.027)	0.644 (-0.019)
w/o Spatial Attention	0.809 (-0.013)	0.658 (-0.005)
w/o Channel Attention	0.805 (-0.017)	0.650 (-0.013)
w/o SPD-Conv	0.814 (-0.008)	0.651 (-0.012)

results corroborate the quantitative improvements and highlight the superiority of M^{2}Former under challenging space environments.

In summary, M^{2}Former achieves a favorable tradeoff between detection performance and model complexity. When combined with AAL and the improved data augmentation strategy, the final model not only surpasses both RT-DETR-R18 and advanced YOLO variants in detection accuracy but also maintains a lightweight and efficient architecture. This demonstrates the practicality and scalability of our design for event-based space object detection.

E. Ablation Study

To validate the effectiveness of each component in the proposed M^{2}Former architecture and the data augmentation strategy, we conduct ablation experiments by individually removing key modules or applying augmentations.

Each module in the M^{2}Former contributes uniquely to detection accuracy. As shown in Table V, removing the Res2Net module results in the most significant performance degradation (−0.027 AP@50 and −0.019 AP@50:95), underscoring its essential role in multiscale feature extraction. Res2Net enhances the receptive field by hierarchically fusing features across channel groups, which is particularly beneficial for resolving fine structures in sparse event data. Eliminating the spatial attention module leads to a noticeable drop (−0.013 AP@50 and −0.005 AP@50:95), confirming its importance in focusing the network on foreground regions and suppressing background noise. The removal of channel attention also leads to accuracy degradation (−0.017 AP@50 and −0.013 AP@50:95). By reweighting feature responses based on global descriptors, this module enhances semantic discrimination and suppresses redundancy. Excluding the SPD-Conv module results in moderate accuracy degradation (−0.008 AP@50 and −0.012 AP@50:95). Unlike standard strided convolutions or pooling, SPD-Conv ensures richer spatial information is retained, benefiting accurate localization.

The quantitative results are further supported by Fig. 6, which visualizes feature activations from the full M^{2}Former and its ablated variants. The baseline model produces spatially precise and semantically coherent activation maps. In contrast, removing Res2Net results in weakened and scattered responses, indicating reduced multiscale sensitivity. Without spatial attention, heatmaps appear diffuse and misaligned, while removing channel attention leads to incomplete and inconsistent activations. The absence of SPD-Conv causes

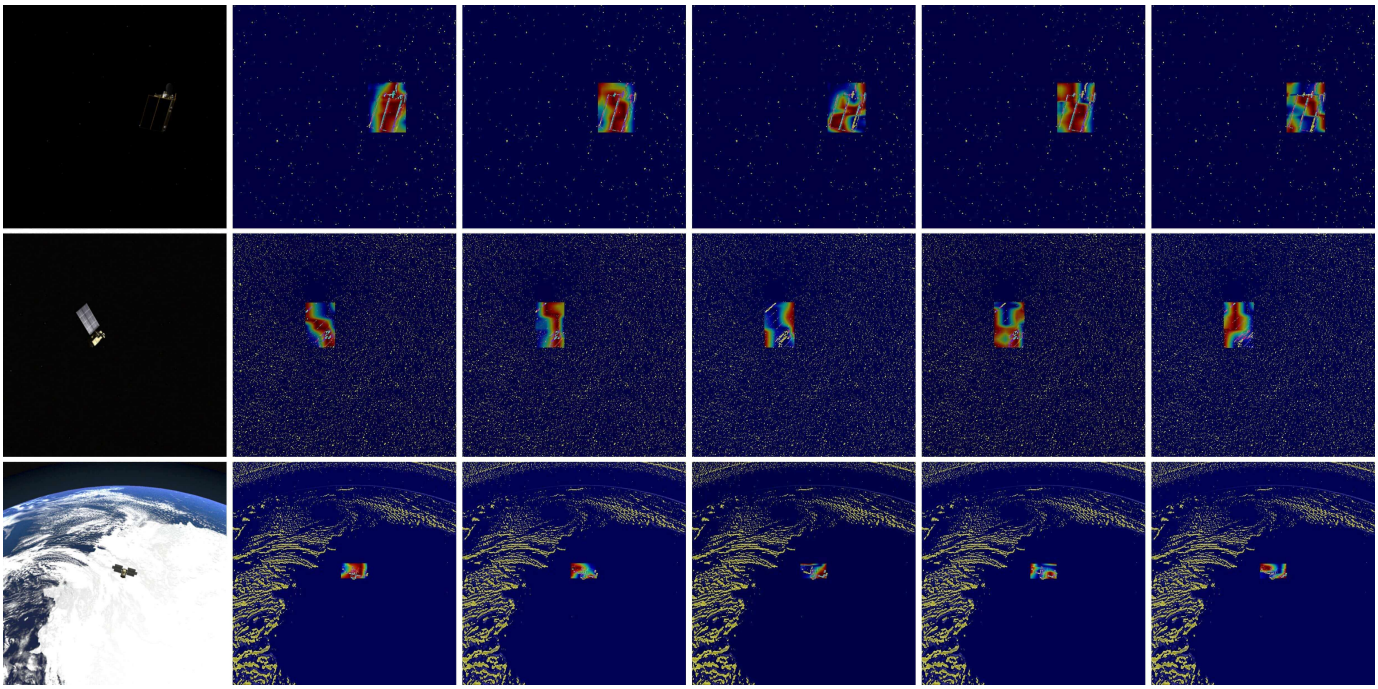

Fig. 6. Heatmap visualizations of M^{2}Former and its ablated variants on event data. Each row shows one test sample. Corresponding RGB reference images and feature maps from the M^{2}Former, and its ablated versions without Res2Net, spatial attention, channel attention, and SPD-Conv are from left to right.

TABLE VI
ABLATION STUDY ON DATA AUGMENTATION

Method	AP@50	AP@50:95
M^{2}Former (baseline)	0.822	0.663
with Mosaic	0.851 (+0.029)	0.698 (+0.026)
with Mixup	0.847 (+0.025)	0.688 (+0.025)
with Transformations	0.837 (+0.015)	0.680 (+0.017)

blurred and spatially distorted responses, likely due to detail loss during downsampling. These patterns visually reinforce the importance of each module in accurate feature localization.

The data augmentation strategy also contributes significantly to model performance. As shown in Table VI, applying Mosaic augmentation on the baseline leads to the largest improvements (+0.029 AP@50 and +0.035 AP@50:95), by increasing sample diversity through multi-image composition. MixUp also yields consistent gains (+0.025 on both metrics) by interpolating samples and labels, enhancing generalization. In comparison, geometric transformations (i.e., translation, scaling, and horizontal flip) provide moderate improvements (+0.015 AP@50 and +0.017 AP@50:95), promoting spatial robustness. These results demonstrate that the augmentations are complementary, and their integration into a unified pipeline is well-justified.

In summary, the ablation study confirms the effectiveness of each component in the M^{2}Former architecture. Their combined use enables efficient multiscale context modeling, attention refinement, and information-preserving downsampling. Additionally, the improved augmentation strategy substantially boosts both detection accuracy and generalization, particularly in the challenging setting of small-object detection in sparse events.

F. Model Robustness and Generalization

To assess the deployment potential of the proposed method, we evaluate its robustness under degraded inputs and its generalization to real event data. Two complementary experiments are conducted: 1) detection performance on low-resolution synthetic event data and 2) zero-shot detection on real event data as well as compared with image modality.

1) Robustness to Lower Resolution: Onboard space systems are often subject to resource constraints and limited bandwidth. To evaluate the model robustness of lower resolution input, we conduct experiments on event resolution of 346 $\times$ 260 (i.e., DAVIS346 camera configuration). Detection results are summarized in Table VII.

Among YOLO variants, YOLOv5m achieves the highest accuracy, and YOLOv8m follows closely, showing that the YOLO family retains strong performance despite degraded resolution. For DETR models, RT-DETR-R18 outperforms M^{2}Former under default settings, primarily due to its larger backbone. Nevertheless, M^{2}Former still delivers superior performance compared with other lightweight backbones. When incorporating AAL, both RT-DETR-R18 and M^{2}Former exhibit consistent improvements, confirming the effectiveness of scale-aware supervision in mitigating the impact of small-object degradation at lower resolution. For data augmentation, however, different behaviors are observed. Aug1 improves RT-DETR-R18 across all metrics, but for M^{2}Former, it unexpectedly reduces APs (0.469 versus 0.488), indicating a decline in small-object detection accuracy. This degradation is likely caused by the additional downscaling in multiscale

TABLE VII
DETECTION PERFORMANCE ON EVENT HISTOGRAM UNDER LOWER RESOLUTION INPUT

Method	AP@50	AP@50:95	APs
YOLOv5s	0.719	0.556	0.414
YOLOv5m	0.796	0.617	0.506
YOLOv8s	0.771	0.598	0.482
YOLOv8m	0.790	0.617	0.503
YOLO11s	0.726	0.564	0.448
YOLO11m	0.728	0.572	0.423
RT-DETR-R18	0.733	0.562	0.492
PoolFormer-S12	0.621	0.448	0.392
RepViT-M0.9	0.659	0.487	0.421
MobileNetV2	0.589	0.434	0.376
M^2Former	0.699	0.524	0.472
RT-DETR-R18 + AAL	0.752	0.580	0.524
RT-DETR-R18 + AAL + Aug1	0.760	0.593	0.526
RT-DETR-R18 + AAL + Aug2	0.785	0.609	0.534
M^2Former + AAL	0.721	0.546	0.488
M^2Former + AAL + Aug1	0.744	0.568	0.469
M^2Former + AAL + Aug2	**0.809**	**0.622**	**0.546**

Note: Underline indicates the best score within each framework, and bold highlights the best score across all methods.

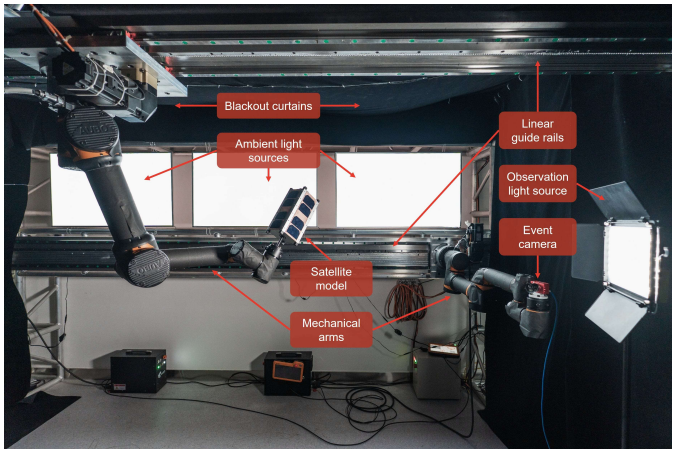

Fig. 7. Simulation testbed for real data collection and its hardware components.

augmentation, which weakens the feature representation of small targets. In contrast, Aug2 brings clear and consistent gains to both frameworks. With AAL and Aug2 combined, M^2Former achieves the best overall performance, surpassing all other configurations.

These results highlight that M^2Former demonstrates strong robustness to lower resolution inputs, the AAL effectively mitigates performance degradation by reinforcing supervision on small objects, and the proposed augmentation strategy further enhances model generalization, avoiding the pitfalls of conventional multiscale augmentation in sparse and noisy event data.

2) Generalization to Real Data: It is difficult to collect a large amount of real data in the context of space object detection. Therefore, to evaluate model generalization, we conduct zero-shot detection on real event data collected from a ground-based testbed using models trained purely on E-SPARK. As illustrated in Fig. 7, the testbed is constructed inside a darkroom equipped with an adjustable lighting system and two motion-controlled mechanical arms to simulate satellite trajectories. A total of 600 samples are collected under two motion settings (normal speed and high speed) and three illumination conditions: normal exposure, overexposure, and underexposure, with 200 samples per condition. Both RGB images and event streams are synchronously captured using a DAVIS346 camera. The satellite targets include CubeSat, Gaofen-13, and the Hubble Space Telescope. For a fair comparison, all models adopt the same training pipeline. We select YOLOv8s as the representative YOLO baseline due to its balance between accuracy and complexity, while RT-DETR-R18 is included as the backbone baseline of our M^2Former. The task is formulated as class-agnostic detection. Experimental results are summarized in Table VIII.

TABLE VIII
ZERO-SHOT DETECTION PERFORMANCE FROM SYNTHETIC DOMAIN TO REAL DOMAIN

Method	Modality	Light	AP@50	AP@50:95
YOLOv8s	RGB	NE	0.014	0.003
		OE	0.001	0.000
		UE	0.000	0.000
		Avg.	0.005	0.001
YOLOv8s	Event	NE	0.184	0.117
		OE	0.143	0.104
		UE	0.258	0.139
		Avg.	0.195	0.120
RT-DETR-R18	Event	NE	0.132	0.059
		OE	0.029	0.009
		UE	0.578	0.286
		Avg.	0.246	0.118
M^2Former	Event	NE	0.332	0.170
		OE	0.137	0.079
		UE	0.624	0.259
		Avg.	0.364	0.169

Note: NE denotes Normal Exposure, OE denotes Overexposure, and UE denotes Underexposure lighting conditions. Avg. denotes the average performance across the three lighting conditions.

While the RGB-based YOLOv8s performs well on synthetic data, it generalizes poorly to real-world RGB inputs, yielding an average AP@50 of just 0.005 and AP@50:95 of 0.001. Notably, it completely fails under underexposed conditions, highlighting the limitations of conventional image-based vision in extreme lighting. In contrast, all event-based models demonstrate significantly better performance. YOLOv8s (event) achieves an average of 0.195 AP@50 and 0.120 AP@50:95. RT-DETR-R18 (event) achieves an average of 0.246 AP@50 and 0.118 AP@50:95. M^2Former (event) achieves the best results across all metrics, reaching 0.446 AP@50 and 0.197 AP@50:95 on average.

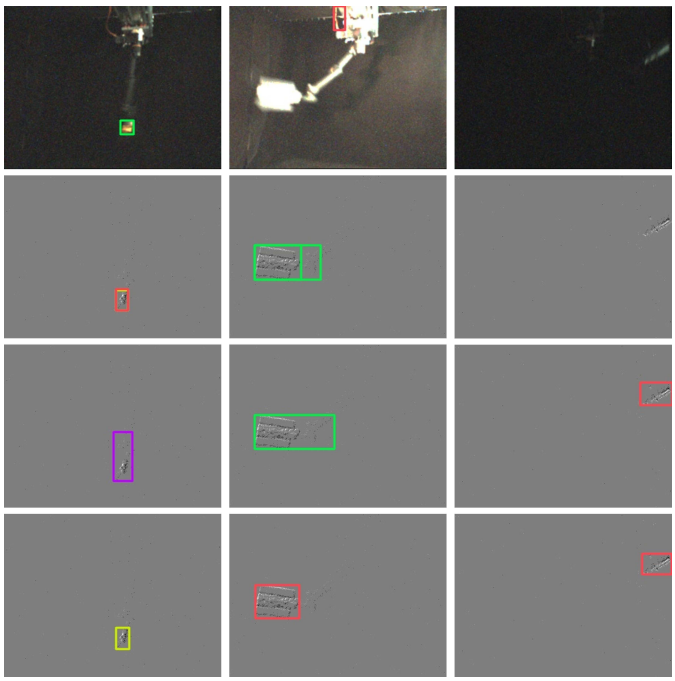

Fig. 8. Detection results under varying illumination conditions. Columns represent normal, overexposed, and underexposed scenes; rows correspond to the results of YOLOv8s (RGB), YOLOv8s (event), RT-DETR-R18 (event), and M²Former (event), respectively. Bounding box colors are defined by the E-SPARK training categories, but they do not carry semantic meaning in the zero-shot detection experiments, which are treated as class-agnostic.

Interestingly, all event-based models achieve their highest performance under underexposed conditions. This phenomenon can be explained by the reflective characteristics of satellite surfaces, which tend to produce glare and unstable responses under strong illumination, thereby introducing noise in event data. In low-light settings, such reflections are suppressed. Combined with the high dynamic range of event cameras, this results in cleaner and more structurally meaningful event signals, ultimately enabling more accurate detection.

We further provide qualitative comparisons in Fig. 8. The YOLOv8s (RGB) struggles to localize the object due to extreme illumination and motion blur. YOLOv8s (event) improves robustness but still suffers from occasional missed detections and redundant boxes due to fixed NMS thresholds. RT-DETR-R18 (event) provides more stable detections, but its output bounding boxes are often coarse and less tightly aligned with the target. In contrast, M²Former (event) provides the most consistent and precise localization across all scenarios. It achieves higher recall and outputs tighter bounding boxes, confirming its superior robustness and generalization.

These results indicate that the event modality is significantly more robust to illumination variation than RGB inputs. Thanks to the NMS-free architecture, RT-DETR-R18 and M²Former outperform YOLOv8s. Although the performance drops noticeably on real event data due to the zero-shot detection setting and the large domain gap between synthetic and real data, M²Former still outperforms the other methods, demonstrating stronger generalization capability.

V. Conclusion

This article presents the first systematic investigation of event-based vision for the challenging task of space object detection. Motivated by the limitations of conventional imaging sensors under extreme illumination and dynamic motion conditions in space, we explore the potential of event cameras, owing to their asynchronous and high-dynamic-range characteristics, to address these challenges effectively.

First, to mitigate the scarcity of real event data, we developed a novel data synthesis pipeline that generates physically consistent event streams from static RGB images through controlled affine transformations and advanced event simulation, resulting in the publicly available E-SPARK dataset.

Next, to enable efficient and accurate detection, we propose M²Former, a lightweight backbone inspired by the MetaFormer paradigm. By integrating efficient convolutional and downsampling operations with multiscale feature extraction capability, M²Former achieves a favorable balance between detection accuracy and computational cost. We further introduce an AAL, which adaptively enhances localization supervision based on object scale. In addition, we utilize an improved data augmentation strategy, incorporating Mosaic, MixUp, and geometric transformations to increase supervision density and data diversity.

Extensive experiments conducted on both synthetic and real event data demonstrate the effectiveness of the proposed method. On synthetic event data, our method achieves the best performance, significantly outperforming RT-DETR-R18 and YOLO models. Under zero-shot detection on real event data, our method maintains robust detection performance across diverse illumination conditions, where RGB-based detectors experience severe degradation. Moreover, M²Former requires over 50% fewer parameters and computations than RT-DETR-R18, making it highly suitable for deployment on resource-constrained spaceborne platforms.

In summary, this work highlights the potential of event-based detectors for space object detection. The proposed method achieves superior detection performance, strong robustness under adverse illumination, and an efficient architecture design suitable for onboard deployment. Future research will focus on developing more effective spatiotemporal representations of event data, optimizing detection frameworks to better exploit their sparse and asynchronous nature, and exploring domain adaptation or simulation-to-real transfer techniques to further improve real-world generalization.

References

[1] Z. Fu et al., "Federated learning-enabled self-reconfigurable satellites for resident space objects detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5621813.

[2] T. Rybus, "Robotic manipulators for in-orbit servicing and active debris removal: Review and comparison," *Prog. Aerosp. Sci.*, vol. 151, Nov. 2024, Art. no. 101055.

[3] J. Rodriguez-Villamizar and T. Schildknecht, "Daylight measurement acquisition of defunct resident space objects combining active and passive electro-optical systems," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625117.

[4] V. C. Coffey, "Seeing in the dark: Defense applications of IR imaging," *Opt. Photon. News*, vol. 22, no. 4, pp. 26–31, 2011.

[5] J. Zhou, "A review of LiDAR sensor technologies for perception in automated driving," *Academic J. Sci. Technol.*, vol. 3, no. 3, pp. 255–261, Nov. 2022.

[6] M. Strube, R. Henry, E. Skeleton, J. V. Eepoel, N. Gill, and R. McKenna, "Raven: An on-orbit relative navigation demonstration using international space station visiting vehicles," in *Proc. AAS GN&C Conf.*, 2015, Paper SFC-E-DAA-TN20551.

[7] S. Jing et al., "ESVT: Event-based streaming vision transformer for challenging object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5607113.

[8] A. Rathinam, H. Qadadri, and D. Aouada, "SPADES: A realistic spacecraft pose estimation dataset using event sensing," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 11760–11766.

[9] Z. Liu, B. Guan, Y. Shang, Y. Bian, P. Sun, and Q. Yu, "Stereo event-based, 6-DOF pose tracking for uncooperative spacecraft," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5607513.

[10] E. Elms, Y. Latif, T. H. Park, and T.-J. Chin, "Event-based structure-from-orbit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 19541–19550.

[11] T.-J. Chin, S. Bagchi, A. Eriksson, and A. Van Schaik, "Star tracking using an event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[12] G. Cohen et al., "Event-based sensing for space situational awareness," *J. Astron. Sci.*, vol. 66, no. 2, pp. 125–141, Jun. 2019.

[13] S. Afshar, A. P. Nicholson, A. van Schaik, and G. Cohen, "Event-based object detection and tracking for space situational awareness," *IEEE Sensors J.*, vol. 20, no. 24, pp. 15117–15132, Dec. 2020.

[14] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, May 2022.

[15] T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal, "Object detection with transformers: A review," 2023, *arXiv:2306.04670*.

[16] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.

[17] T. H. Park, M. Märtens, G. Lecuyer, D. Izzo, and S. D'Amico, "SPEED+: Next-generation dataset for spacecraft pose estimation across domain gap," in *Proc. IEEE Aerosp. Conf. (AERO)*, Mar. 2022, pp. 1–15.

[18] P. F. Proença and Y. Gao, "Deep learning for spacecraft pose estimation from photorealistic rendering," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Paris, France, May 2020, pp. 6007–6013.

[19] M. Bechini, M. Lavagna, and P. Lunghi, "Dataset generation and validation for spacecraft pose estimation via monocular images processing," *Acta Astronautica*, vol. 204, pp. 358–369, Mar. 2023.

[20] H. A. Dung, B. Chen, and T.-J. Chin, "A spacecraft dataset for detection, segmentation and parts recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2012–2019.

[21] M. A. Musallam et al., "Spacecraft recognition leveraging knowledge of space environment: Simulator, dataset, competition design and analysis," in *Proc. IEEE Int. Conf. Image Process. Challenges (ICIPC)*, Sep. 2021, pp. 11–15.

[22] Y. Cao, J. Mu, X. Cheng, and F. Liu, "Spacecraft-DS: A spacecraft dataset for key components detection and segmentation via hardware-in-the-loop capture," *IEEE Sensors J.*, vol. 24, no. 4, pp. 5347–5358, Feb. 2024.

[23] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[24] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[25] R. Khanam and M. Hussain, "What is YOLOv5: A deep look into the internal features of the popular object detector," 2024, *arXiv:2407.20892*.

[26] R. Varghese and M. Sambath, "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *Proc. Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, Apr. 2024, pp. 1–6.

[27] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.

[28] N. AlDahoul, H. A. Karim, A. De Castro, and M. J. T. Tan, "Localization and classification of space objects using EfficientDet detector for space situational awareness," *Sci. Rep.*, vol. 12, no. 1, p. 21896, Dec. 2022.

[29] N. Shen, R. Xv, Y. Gao, C. Qian, and Q. Chen, "An improved YOLOv5 model based on feature fusion and attention mechanism for multiscale satellite recognition," *IEEE Sensors J.*, vol. 24, no. 12, pp. 19385–19396, Jun. 2024.

[30] H. Zhang, Y. Zhang, Q. Feng, and K. Zhang, "A semi-supervised object detection method for close range detection of spacecraft and space debris," *Int. J. Aeronaut. Space Sci.*, vol. 26, no. 2, pp. 773–784, Mar. 2025.

[31] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Jun. 2006, pp. 850–855.

[32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[33] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[34] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[35] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," 2022, *arXiv:2201.12329*.

[36] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query DeNoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13619–13627.

[37] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[38] W. Yu et al., "MetaFormer baselines for vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 896–912, Feb. 2024.

[39] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10819–10829.

[40] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Rep ViT: Revisiting mobile CNN from ViT perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15909–15920.

[41] L. Cordone, B. Miramond, and P. Thierion, "Object detection with spiking neural networks on automotive event data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[42] J. Zhang et al., "Spiking transformers for event-based single object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 8801–8810.

[43] Q. Su et al., "Deep directly-trained spiking neural networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6532–6542.

[44] Y. Li et al., "Graph-based asynchronous event processing for rapid object recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 934–943.

[45] S. Schaefer, D. Gehrig, and D. Scaramuzza, "AEGNN: Asynchronous event-based graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12371–12381.

[46] D. Gehrig and D. Scaramuzza, "Low-latency automotive vision with event cameras," *Nature*, vol. 629, no. 8014, pp. 1034–1040, May 2024.

[47] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5419–5427.

[48] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1346–1359, Jul. 2017.

[49] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 989–997.

[50] A. Tomy, A. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, "Fusing event-based and RGB camera for robust object detection in adverse conditions," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 933–939.

[51] J. Li, J. Li, L. Zhu, X. Xiang, T. Huang, and Y. Tian, "Asynchronous spatio-temporal memory network for continuous event-based object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2975–2987, 2022.

[52] D. Li, Y. Tian, and J. Li, "SODFormer: Streaming object detection with transformer using events and frames," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 14020–14037, Nov. 2023.

[53] Y. Peng, H. Li, Y. Zhang, X. Sun, and F. Wu, "Scene adaptive sparse transformer for event-based object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16794–16804.

[54] E. Pérot, P. D. Tournemire, D. O. Nitti, J. Masci, and A. Sironi, "Learning to detect objects with a 1 megapixel event camera," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 16639–16652.
[55] M. Gehrig and D. Scaramuzza, "Recurrent vision transformers for object detection with event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 13884–13893.
[56] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor, "Converting static image datasets to spiking neuromorphic datasets using saccades," *Frontiers Neurosci.*, vol. 9, p. 437, Nov. 2015.
[57] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3D perception," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2032–2039, Jul. 2018.
[58] P. De Tournemire, D. Nitti, E. Perot, D. Migliore, and A. Sironi, "A large scale event-based detection dataset for automotive," 2020, *arXiv:2001.08499*.
[59] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. 2nd Conf. Robot Learn. (CoRL)*, 2018, pp. 969–982.
[60] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.
[61] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[62] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.
[63] A. Z. Zhu, Z. Wang, K. Khant, and K. Daniilidis, "EventGAN: Leveraging large scale image datasets for event cameras," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2021, pp. 1–11.
[64] N. Messikommer, D. Gehrig, M. Gehrig, and D. Scaramuzza, "Bridging the gap between events and frames through unsupervised domain adaptation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3515–3522, Apr. 2022.
[65] D. Jian and M. Rostami, "Unsupervised domain adaptation for training event-based networks using contrastive learning and uncorrelated conditioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18675–18685.
[66] J. Cao, X. Zheng, Y. Lyu, J. Wang, R. Xu, and L. Wang, "Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 9026–9032.
[67] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[68] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[69] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.
[70] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, Sep. 2022, pp. 443–459.
[71] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "VarifocalNet: An IoU-aware dense object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8514–8523.
[72] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
[73] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.
[74] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "DEIM: DETR with improved matching for fast convergence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 15162–15171.
[75] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

Ruitao Pan received the B.E. degree in intelligent manufacturing engineering from Shantou University, Shantou, Guangdong, China, in 2022. He is currently pursuing the Ph.D. degree in control science and engineering with the School of Future Technology, Xi'an Jiaotong University, Xi'an, China.

His research interests include event-based vision and 3-D vision.

Chenxi Wang received the Ph.D. degree in mechanical engineering from X'ian Jiaotong University, Xi'an, China, in 2020.

He is currently an Associate Professor with the School of Mechanical Engineering, Xi'an Jiaotong University. His research interests include rocket engine status monitoring and fault diagnosis, in-orbit spacecraft threat identification and trajectory prediction, satellite formation dynamics modeling and orbit–attitude control, in-orbit antenna assembly and manufacturing, space 3-D printing, helicopter system identification and active vibration control, additive manufacturing process parameter optimization, process quality monitoring, and intelligent feedback control.

Bin Han (Student Member, IEEE) received the B.S. degree from Wuhan University of Technology, Wuhan, Hubei, China, in 2022. He is currently pursuing the Ph.D. degree in mechanical engineering with the School of Future Technology, Xi'an Jiaotong University, Xi'an, China.

His research interests include spacecraft pose estimation and deep learning.

Xinyu Zhang received the B.E. degree in mechatronics engineering from Northeast Forestry University, Harbin, Heilongjiang, China, in 2023. She is currently pursuing the M.E. degree in mechanical engineering with the School of Future Technology, Xi'an Jiaotong University, Xi'an, China.

Her research interests include spacecraft identification and intelligent perception for non-cooperative targets.

Zhi Zhai received the B.S. degree in mechanical manufacturing and automation and the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2007 and 2014, respectively.

She is currently a Researcher with the School of Mechanical Engineering, Xi'an Jiaotong University. She has a background in prognosis and health monitoring of spacecraft. Her research interests include remaining useful life (RUL) prediction of satellite battery, and fault diagnosis of liquid rocket engine.

Jinxin Liu received the B.S. degree in mechanical engineering from the School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 2011, and the Ph.D. degree in mechanical engineering from the School of Mechanical Engineering, X'ian Jiaotong University, in 2016.

He was a Visiting Scholar with the Lawrence Berkeley National Laboratory, Berkeley, CA, USA, from 2014 to 2015. He is currently a Professor with the School of Mechanical Engineering, Xi'an Jiaotong University. He is working on active safety control of machinery and equipment, including active vibration control, fault-tolerant control, and precision control of large-scale, complex, and dynamical systems.

Naijin Liu received the B.S. degree in electronic information engineering from the University of Science and Technology of China, Hefei, China, in 2002, and the Ph.D. degree in communication and information systems from the University of Science and Technology of China, in 2007.

He is currently the Deputy Director of the Qian Xuesen Space Technology Laboratory, China Academy of Space Technology, Beijing, China. He is also a Professor with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China. His research interests include satellite communication and space intelligent information networking.

Dr. Liu is a Council Member of the Chinese Society of Astronautics.

Xuefeng Chen (Senior Member, IEEE) received the Ph.D. degree from the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China, in 2004.

He is currently a Full Professor with the School of Mechanical Engineering, Xi'an Jiaotong University. He hosted the National Key 973 Research Program of China as a Principal Scientist in 2015. His research interests include finite-element method, mechanical signal processing, and fault diagnosis.

Dr. Chen is a member of American Society of Mechanical Engineers (ASME). He received the National Excellent Doctoral Thesis Award in 2007, the First Technological Invention Award of Ministry of Education in 2008, the Second National Technological Invention Award in 2009, the First Provincial Teaching Achievement Award in 2013, the First Technological Invention Award of Ministry of Education in 2015, and the Science and Technology Award for Chinese Youth in 2013. He is the Chair of the IEEE Xi'an and Chengdu Joint Section Instrumentation and Measurement Society Chapter. He works as the Executive Director of the Fault Diagnosis Branch, China Mechanical Engineering Society.