VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows

Xiao Wang, Member, IEEE, Jianing Li, Member, IEEE, Lin Zhu, Member, IEEE, Zhipeng Zhang, Fellow, IEEE, Zhe Chen, Member, IEEE, Xin Li, Yaowei Wang, Member, IEEE, Yonghong Tian, Fellow, IEEE, and Feng Wu, Fellow, IEEE

Abstract—Different from visible cameras which record intensity images frame by frame, the biologically inspired event camera produces a stream of asynchronous and sparse events with much lower latency. In practice, visible cameras can better perceive texture details and slow motion, while event cameras can be free from motion blurs and have a larger dynamic range which enables them to work well under fast motion and low illumination (LI). Therefore, the two sensors can cooperate with each other to achieve more reliable object tracking. In this work, we propose a large-scale Visible-Event benchmark (termed VisEvent) due to the lack of a realistic and scaled dataset for this task. Our dataset consists of 820 video pairs captured under LI, high speed, and background clutter scenarios, and it is divided into a training and a testing subset, each of which contains 500 and 320 videos, respectively. Based on VisEvent, we transform the event flows into event images and construct more than 30 baseline methods by extending current single-modality trackers into dual-modality versions. More importantly, we further build a simple but effective tracking algorithm by proposing a cross-modality transformer, to achieve more effective feature fusion between visible and event data. Extensive experiments on the proposed VisEvent dataset, FE108, COESOT, and two simulated datasets (i.e., OTB-DVS and VOT-DVS), validated the effectiveness of our model. The dataset and source code have been released on: https://github.com/wangxiao5791509/VisEvent_SOT_Benchmark.

Index Terms—Dynamic vision sensors (DVSs), event camera, neuromorphic vision, self-attention and transformers, visual tracking.

I. INTRODUCTION

VISUAL tracking aims at locating the initialized object in the first frame with a bounding box and adjusting the box to better fit the target object for subsequent video frames. It has been widely used in many applications, including intelligent video surveillance, robotics, and autonomous vehicles. With the help of deep learning, many representative deep trackers are proposed [1], [2], [3], [4], [5], [6], [7], [8], [9]. To be specific, Dong et al. [5] found that the samples given in the first frame (termed decisive samples) are ignored during offline training and propose a compact latent network that can quickly adjust the tracking model to work well in new scenarios. Shen et al. [6] proposed a teacher–student knowledge distillation model that supports the learning of a small but fast tracker from large Siamese trackers. Dong and Shen [7] also exploited a new triplet loss function to enhance the deep features for Siamese-based tracking. Different from existing Siamese trackers which adopt depthwise cross-correlation (DW-XCorr) for activation response maps prediction, Han et al. [8] proposed the asymmetric convolution (ACM) operators for tracking which is a learnable and flexible module. However, due to the utilization of RGB cameras, existing trackers still suffer from challenging scenarios, such as low illumination (LI), fast motion (FM), and background clutter.

To handle aforementioned challenges, some researchers resort to biologically inspired event cameras like dynamic vision sensors (DVSs) [10] for target object tracking [11], [12], [13], [14], [15]. Different from regular visible cameras which record an intensity image in a frame manner (high latency, i.e., 10–20 ms), the event cameras output a stream of asynchronous events. The pixels of event cameras send information independently only when visual intensity changes (also called an event). Therefore, the event sensors excel at capturing the motion information with very low latency (1 μs) and are almost free from the trouble of motion blur (MB). It also


Fig. 1. Illustration of complementary characteristics of Visible and Event cameras. From subfigure (a)–(c), we can find that the Event camera works well in LI, high speed, and even in clutter background scenarios due to its advantages in high dynamic range, dense temporal resolution, and unique imaging features. In contrast, the visible camera performs poorly when facing these challenging attributes, however, it works well in slow motion or static scenarios and is good at capturing the color and texture information according to subfigure (d). Therefore, a more reliable and accurate tracking result can be obtained if we combine the two sensors.

Fig. 2. Sampling mechanisms of visible and event cameras. Each pixel in the visible image records the light intensity (I) in a synchronous way; while each pixel in the biologically inspired event camera asynchronously reflects the changes in lighting intensity. Usually, we use $I_{-1}$ to denote the ON/OFF event (enhancement/diminished light). $V_d$ is the neuronal membrane potential.

requires much less energy, bandwidth, and computation. In addition, DVS sensors also outperform the visible cameras on dynamic range (140 versus 60 dB), which enables them to work effectively even in poor illumination conditions. The advantages of the latency, resource consumption, and operation environments make the event cameras more suitable for target tracking in challenging scenarios. The comparison of the imaging quality and sampling mechanism of the two sensors are given in Figs. 1 and 2, respectively, to help readers better understand their unique advantages.

Despite benefits, the event cameras cannot capture slow-motion or static objects and lack fine-grained texture information which is also very important for high-performance tracking. Therefore, the integration of visible cameras and DVS sensors is an intuitive idea for reliable object tracking. There are already several works [14], [15], [16], [17] developed based on this setting, but their experiments are conducted on simulation data or several simple videos. Their performance on real data in the wild is still unknown. Recently, Zhang et al. [18] proposed a new dataset that contains 108 videos, termed FE108, but tracking performance on this dataset is almost saturated. The development of this field is rather slow compared with visible camera-based tracking due to the lack of a large-scale Visible-Event-based object tracking dataset.

In this work, we first propose a large-scale neuromorphic tracking benchmark that contains 820 Visible and Event video sequence pairs, termed VisEvent. This dataset fully reflects the challenging factors in the real world, such as MB, fast and slow motion, LI, high dynamic range, background clutter, etc. It contains 17 attributes and mainly focuses on traffic scenes, thus the target objects are mainly people and vehicles. To construct a comprehensive benchmark, we also recorded some videos from the indoor scene. In total, our dataset contains 371 128 frames. We split them into the training and testing subsets, each of them containing 500 and 320 video pairs, respectively. Due to the lack of baseline methods to be compared for future works, we extend currently visible camera-based trackers into dual-modality versions with different fusion strategies, such as early, middle, and late fusion.

On the basis of our newly proposed VisEvent dataset, we further build a novel and effective baseline method by developing a cross-modality transformer (CMT) module. The proposal of our CMT module is based on the following two observations and key insights.

1) Existing trackers [19], [20], [21] usually adopt convolutional neural networks for tracking, which only learn the local features well using limited convolutional kernels. The most recent RGB trackers also exploit self-attention or Transformers networks for global representation learning [22], [23].

2) Different from existing Transformer-based trackers, we need to consider the modality interactions between RGB and Event streams to achieve a more robust feature fusion.

Therefore, a cross-attention layer is proposed to connect the dual modalities for interactive message passing. Then, a self-attention layer is adopted to learn and enhance the feature representations in a global view. Thus, our proposed CMT boosts the inter- and intra-modality features and significantly improves the final tracking results. As validated in our experiments, it can be integrated into existing binary classification trackers, such as MDNet [19], RT-MDNet [20], and discriminative correlation filter (CF) trackers ATOM [21]. More details can be found in Fig. 3 and Section III-C.

Generally speaking, the contributions of this article can be concluded as the following three aspects.

1) We introduce a comprehensive neuromorphic tracking dataset comprising 820 Visible-Event videos, termed VisEvent. This marks the inception of a large-scale Visible-Event benchmark dataset collected from real-world scenarios, tailored specifically for single object tracking.

2) We present a straightforward yet highly effective baseline tracker achieved through the development of a CMT

The video tutorial of this work can be found at https://youtu.be/vGwHI2d2AX0.
module. This module adeptly leverages the distinctive characteristics of various modalities to enhance tracking robustness. Notably, this is the first instance where the successful application of a CMT in visible-event tracking has been demonstrated.

3) We have assembled a diverse set of more than 35 dual-modality-based trackers for our benchmark dataset. These trackers serve as valuable resources for future research, enabling comprehensive comparisons across various tracking pipelines (e.g., CF-based, binary classification-based, and Siamese matching-based trackers) and fusion strategies (e.g., early, middle, and late fusion).

II. RELATED WORK

Visible Camera-Based Tracking: Most of the current trackers are developed based on RGB cameras and track the target object frame by frame. Traditional RGB trackers use hand-crafted features for target representation but perform poorly in challenging scenarios. Among them, CF-based trackers dominate the tracking field due to their high efficiency and good performance [24], [25], [26]. After that, the deep learning trackers, especially the Siamese network-based trackers began to occupy the top positions of various benchmarks. Shen et al. [27] proposed an attention-based Siamese network which can improve matching performance by a sub-Siamese network. The feature learning and classification capabilities of extreme learning machine (ELM) are exploited by Deng et al. [28] for efficient visual tracking. Liu et al. [29] attempted to address the occlusion issue in the tracking task using correlation filtering and probabilistic finite state machines (FSMs). Zhou et al. [30] proposed a gradient-guided feature adjustment module to generate target-aware features for constructing the state estimation network, which achieves high-performance visual tracking. Li et al. [31] developed a dual-regression tracking framework by combining the discriminative fully convolutional module and a fine-grained CF component. Li et al. [32] treated the TIR tracking as a similarity verification task, and propose a Hierarchical Spatial-aware Siamese CNN (named HSSNet) for TIR tracking. Dong et al. [33] exploited the parameter optimization in tracking task using deep reinforcement learning and achieve a higher tracking performance. Lu et al. [34] proposed a new shrinkage loss to handle the data imbalance issue when learning deep features for tracking. Dong et al. [35] argued that the training instances are ignored in existing trackers and the authors propose a new quadruplet deep network that obtains more powerful features for single object tracking. Liang and Shen [36] proposed the Local Semantic Siamese (LSSiam) network which not only learns global features well but also local semantic features (which contains more fine-grained and partial information) for visual tracking. Cao et al. [37] proposed an autonomous underwater vehicle tracking control algorithm that handles the underwater dynamic target tracking task well by predicting its trajectory. Choi et al. [38] addressed the task of target tracking on the sphere by considering topographic structure.

Recently, the Transformer networks are widely exploited in visual tracking task and achieve higher performance on multiple benchmark datasets [23]. However, their performance under LI, FM, and low resolution is still unsatisfactory. Many works are proposed to handle these issues, such as active hard sample generation [39], [40] and deblur [41], however, the existing algorithm cannot address these issues well due to the bad imaging quality of RGB cameras. Other sensors are also explored for tracking task, including high frame rate cameras (short for HFR, larger than 200 FPS), thermal cameras, and depth cameras, but HFR cameras are sensitive to illumination, thermal cameras are expensive, and depth cameras are also helpless for high speed and low light, which limit their wide applications in practical scenarios.

Event Camera-Based Tracking: Compared with RGB trackers, few people pay attention to tracking based on event cameras. Chen et al. [11], [14] proposed the adaptive timesurface with linear time decay (ATSLTD) event-to-frame conversion algorithm for event frame construction and redetect the target object when model drifting. The synchronous
time-surface with linear time decay (TSLTD) representation is explored and fed into a CNN-LSTM network for 5-DoF object motion regression in [11]. To handle the issue of local search in event-based tracking, Ramesh et al. [42] proposed a data-driven, global sliding window-based detector to help redetect the target object when it re-enters the field-of-view of the camera. Hernandez et al. [43] and Lopez and Chli [44] explored the high-speed feature tracking with DVS sensors. Cao et al. [45] proposed a target tracking controller based on a spiking neural network that can be deployed on autonomous robots. Jiang et al. [46] also proposed a tracking framework that contains an offline-trained detector and an online-trained tracker which complement each other. Zhu et al. [47] proposed a density-insensitive downsampling strategy to get the key events and employ graph neural networks to capture the spatiotemporal cues for event-based tracking. Zhang et al. [48] exploited the cross-style and cross-frame-rate alignment between the visual and event data for accurate tracking. Although these trackers work well in simple scenarios, however, their performance on large-scale tracking benchmarks is still unknown. Also, the performance of these models on tracking objects that rarely move or are stationary is still alarming.

**Tracking by Combining Visible and Event Cameras:** Joint utilizing the two sensors for robust tracking is an intuitive idea and the initial verification has been obtained in the following work. For example, [49] and [50] first propose asynchronous photometric feature tracking with the event and RGB sensors. Liu et al. [16] also attempted to extract candidate ROIs from RGB frames and event flows simultaneously for more accurate tracking. Huang et al. [17] developed an SVM-based tracker using reconstructed samples for an online update and candidate search locations mining from event flows with a CeleX sensor. DashNet [15], [51] is developed based on parallel SNN and CNN tracking and fusion which can run at 2083 FPS on neuromorphic chips. Their work fully demonstrates the vast potential of Visible-Event tracking in practical applications. Tang et al. [52] proposed a single-stream multimodal tracking framework based on the Transformer network which directly takes the RGB frame and event voxel as input.

These works have made preliminary explorations in this direction, however, their experiments are conducted on several simple real videos or simulation data, as shown in Table I. Their results on really challenging scenarios are still unknown, also, their work lacks proper baseline methods to compare. We believe our proposed dataset and baseline algorithm will be a good platform for research in this direction.

### III. METHODOLOGY

#### A. Motivation and Overview

Based on our proposed VisEvent dataset, we first extend current trackers which are developed for RGB videos into dual-modality versions and evaluate their results. According to the experimental results, we observe that existing trackers are less effective on our dataset even though deep neural networks and regular attention modules are used. For instance, as shown in Fig. 7, the SiamRPN++ [53] and SuperDiMP [54] only attains 0.576(0.410 and 0.489)0.320 on precision and success plot, respectively. MDNet [19] with channel and spatial attention (SAtt) only achieves 0.456(0.273 and 0.455)0.270, as listed in Table VIII. How to design a more effective information fusion module for visible-event tracking is still a question worth exploring.

In this article, we propose a new tracking algorithm by fully exploiting the RGB frame and event stream. The key motivation of our tracker is existing trackers usually adopt convolutional neural networks for tracking, which only learn the local features well using limited convolutional kernels. The most recent RGB/Event trackers also exploit self-attention or Transformer networks for global representation learning [23], [55]. On the other hand, the modality interactions between RGB and event stream need to be considered for high-performance tracking. Given the input modalities, we simply use the event images transformed from event flows to fuse with visible images. The convolutional neural network is used for feature extraction, more importantly, a simple but effective feature fusion module is proposed to achieve interactive learning via information propagation between dual modalities, termed CMTs. As shown in Fig. 3, we first extract the feature representations of visible and event images using CNN. Then, the candidate proposals are sampled and fed into the RoI align module for instance feature extraction. Afterward, we attain the base vector \( m \) with element-wise multiplication operations based on the given feature representations of dual modalities. The base vector is used as the query vector to attend the two modalities (context vector), respectively, to realize the interaction and transmission of information flows. After that, we use self-attention layers to boost the internal connections of each modality. Finally, the two features are concatenated and fed into the classifier for tracking.

#### B. Input Representation

Given the synchronous video frames and asynchronous event flows, how to represent and adaptively fuse the two modalities is the key to successful and reliable visual tracking on the VisEvent. From the perspective of the perception principle, the visible cameras capture the global scene by recording the intensity of all pixels in a frame manner. Usually, the CNN is used to extract its feature representations, for example, the RT-MDNet [20] uses three convolutional layers as its backbone network. Different from visible sensors, the event cameras asynchronously capture the variation in log-scale intensity \( I \), in other words, each pixel will output a discrete event independently when the visual change exceeds a threshold \( \theta \)

\[
||\log(I_{t+1}) - \log(I_t)|| \geq \theta. \tag{1}
\]

In practice, we use a 4-tuple \( \{x, y, t, p\} \) to represent the discrete event of a pixel captured with DVS, where the \( x \) and \( y \) are spatial coordinates, \( t \) is the timestamp, and \( p \) is the polarity of brightness variation, that is, 1 and \( -1 \) are used to denote the ON event (increase) and OFF event (decrease), respectively. A comparison of the sampling mechanism of visible and event cameras is visualized in Fig. 2.
To fully utilize the benefits of CNN, previous event trackers [11], [14], [42] usually transform the asynchronous event flows into synchronous event image by stacking the events in a fixed time interval. In this work, we also adopt such transformation to get the event images but focus on designing novel feature fusion modules for high-performance tracking. In the subsequent section, we will introduce our proposed CMT for interactive dual-modal information propagation.

C. Cross-Modality Transformer for Fusion

Following RT-MDNet [20], we take the three convolutional layers as the shared backbone of our tracker. Once we obtain the feature representation of dual modalities, we can directly fuse them for tracking. However, the dual features are extracted independently and thus lack interactive feature learning which may limit its representation ability. Many works demonstrate that joint feature learning between multimodal data will bring more powerful feature representation. Inspired by previous works [56], [57], the CMT is proposed in this work to enhance the message passing between dual modalities. This module is developed based on an attention mechanism that targets at retrieving information from context vectors \( y_j \) based on query \( x \). Usually, we can first compute the similarity score \( a_j \) between the query \( x \) and context vector \( y_j \) using MLP layers. Then, this score will be normalized with the Softmax operator. Finally, the context vectors will be weighted and summed as the output of an attention layer: 
\[
\text{Att}_{x ightarrow y}(x, \{y_j\}) = \sum_j a_j y_j.
\]

The widely used self-attention layer [58] is a special case of attention family, as its query vector \( x \) is actually from the context vectors.

In our scenario, we have dual-modality features that can be used to attend to each other using a cross-attention model, that is, from RGB to event, and from event to RGB. As shown in Fig. 3, the features of RGB frame and event flows are first fed into a cross-attention module to guide the information propagation along both directions. Formally, we use \( F_e \) and \( F_v \) to denote the initial feature obtained from the CNN backbone network. Then, these two feature maps are added along the channel dimension and reshaped into feature vectors \( \tilde{F}_e \) and \( \tilde{F}_v \). A base vector \( m \) can be attained by element-wise product between input features of dual-modalities, that is, \( m = \tilde{F}_v \odot \tilde{F}_e \). The base vector \( m \) is used as the query feature to attend the context vectors, that is, visible and event features, respectively

\[
\tilde{F}_e = \text{CrossAtt}_{V \rightarrow E}(m, F_v) \quad \tilde{F}_v = \text{CrossAtt}_{E \rightarrow V}(m, F_e).
\]

Therefore, the joint cross-modality representations can be attained with a cross-attention model which can align the dual modalities by exchanging the information.

To boost the internal connections, we introduce the self-attention layers [58] based on the output of the cross-attention model
\[
\mathcal{F}_e = \text{SelfAtt}_{E \rightarrow E}(\tilde{F}_e, \tilde{F}_e), \quad \mathcal{F}_v = \text{SelfAtt}_{V \rightarrow V}(\tilde{F}_v, \tilde{F}_v).
\]

For simplicity, we take the event feature \( \tilde{F}_e \) as an example, and similar operations are implemented for visible features. Specifically speaking, we first use two FC layers with weights \( W_h \) and \( W_g \) to process the input \( \tilde{F}_e \) separately. The output will be fed into a Softmax layer and multiplied with the results of another branch (i.e., \( W_o \odot \tilde{F}_e \)). Finally, we feed these results into an FC layer with weights \( W_p \) to get the attended event features. The aforementioned process can be summarized as follows:
\[
\mathcal{F}_e = W_p(\text{Softmax}((W_h \ast \tilde{F}_e) \ast (W_g \ast \tilde{F}_e)) \ast (W_o \ast \tilde{F}_e))
\]

D. Training and Tracking Phase

In this work, we follow a binary classification-based tracking framework [20] and conduct tracking by discriminating whether the given proposal is a target object or not. In the training phase, we introduce a set of domain-specific layers (DS layers) for each video sequence to learn the shared features that are only used in the training phase. The binary cross-entropy loss \( \mathcal{L}_{ce} \) and instance embedding loss \( \mathcal{L}_{ie} \) are used for the optimization of our network for the RT-MDNet, that is, \( \mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{ie} \). To be specific, the \( \mathcal{L}_{ce} \) can be formulated as follows:
\[
\mathcal{L}_{ce} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{c=1}^{2} \log \left( \sigma_{cl}(\tilde{f}_j(b))_{c\delta(b)} \right)
\]

where \( \tilde{y}_j \) is the ground truth. The value of \( [\tilde{y}_j]_{c\delta} \) is one when the class of a bounding box in the domain \( d \) is \( c \). \( b \) denotes the iterate index. The predicted score of \( j \)th proposal is \( \tilde{f}_j \). For the instance embedding loss \( \mathcal{L}_{ie} \)
\[
\mathcal{L}_{ie} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{d=1}^{D} \log \left( \sigma_{inst}(f_j')_{+d} \right)
\]

Here, the number of domains is \( D \). We refer the readers to check their paper for the details of the two-loss functions [20].
In the testing phase, we train an online classifier using samples extracted from the first frame. For the subsequent frames, we extract proposals with the Gaussian sampling method around tracking result of the previous frame, then, feed them into the classifier to get the response score. The proposal with the maximum score will be chosen as the tracking result of the current frame. In addition, we also use hard sample mining and online update strategy for better tracking.

E. Implementation Details

For our baseline, we first extend MDNet/RT-MDNet into dual-modality and train it on VisEvent dataset for 50 epochs. The learning rate is 0.0001, batch size is 8, and other parameters are default. The training costs about 3 h. For the first frame, we extract 500 positive and 5000 negative samples and train an online classifier for 50 iterations. For other trackers, the LF and MF-based trackers are trained on VisEvent with its default settings. We adopt the pretrained models of EF-based trackers for the testing. All extended trackers have been released to help researchers reproduce our experiments.2

IV. VisEvent Benchmark Dataset

A. Protocols

The VisEvent is developed to provide a dedicated platform for the training and evaluation of Visible-Event tracking algorithms. Therefore, we obey the following protocols when constructing our benchmark.

1) Large Scale: It is important to provide a huge amount of video sequences for data-hungry deep trackers. We collect 820 video pairs with an average of 450 frames for each video.
2) High-Quality Dense Annotations: Our dataset is densely annotated for each frame and is independently checked by a professional labeling company and two PhDs.
3) Short-Term and Long-Term Tracking: Our dataset contains 709 and 111 videos for short-term and long-term tracking which will be beneficial for constructing a robust and flexible tracker.
4) Long-Tail Distribution: In our real world, pedestrians, and vehicles are more related to our life and the two categories occupy the majority of our videos.
5) The Balance Between Dual Modalities: Since the VisEvent contains two modalities, the balance of difficult videos for each modality is very important. The cases where tracking with a single modality can already realize high performance should be avoided.
6) Comprehensive Baselines: We construct multiple baselines for future work to compare by extending visible trackers into their dual-modality version with various fusion strategies. Also, we propose a simple but effective CMT-based tracker as our advanced baseline approach.

B. Data Collection and Annotation

Based on the aforementioned protocols, we first collect multiple video sequences with DVSs, as shown in Fig. 4, which can output visible video frames and event flows simultaneously. It is worth noting that two streams are generated from a single sensor and are already aligned by the hardware. Therefore, no external processing operations like registration in spatial and temporal views are needed. The resolution of dual modalities is 346 × 260. The target objects are UAV, Hand, Pen, Bottle, Tank, Toy, Car, Tennis, Pedestrian, Badminton, Basketball, Book, Plant, Shoes, Phone, Laptop, Bag, and Cat. Parts of them are visualized in Figs. 1 and 6, and more samples can be found in our demo videos.

After acquiring these videos, we first transform the output file format *.aedat4 into RGB and event frames *.bmp and then select video clips that contain a consistent target object as one sequence. The annotation for each frame is fulfilled by a professional label company and two authors of this work checked all the annotations frame by frame. Rough annotations will be adjusted again to further ensure the accuracy of our dataset. Some samples of our dataset are visualized in Fig. 6.

C. Attribute Definition

As shown in Table II, there is a total of 17 attributes defined in our proposed VisEvent dataset. For the RGB cameras, our dataset reflects many popular attributes in single object tracking, such as camera motion, rotation, scale variation, and occlusion. For the event cameras, since it is challenging to track the static target object, we introduce the attribute NO motion (NOM) to evaluate the tracking performance under this situation. Other motion-related challenges are also considered, including FM, MB, and background object motion (BOM). Our dataset also reflects the scenarios under different lighting conditions, such as LI, over exposure (OE), and illumination variation (IV). It is worth noting that all our videos are with low resolution (i.e., 346 × 260) compared with resolution 1280 × 720 in GOT-10K [61] due to the limitation of the hardware. Therefore, we do not explicitly list low-resolution attributes in Table II. The distribution of each attribute in our dataset will be presented in subsequent sections and visualized in Fig. 5.

D. Statistical Analysis

As shown in Fig. 5 (left subfigure), our proposed VisEvent tracking dataset contains 820 video sequence pairs (371 RGB frames total), the minimum, maximum, and average length are 18, 6246, and 450 frames, respectively. The frame rate of visible videos is about 25 FPS. For the distribution of video length, we have 276, 222, 76, 65, 75, and 111 videos for

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE II

DESCRIPTION OF 17 ATTRIBUTES IN OUR VISEVENT DATASET

| Attributes | Definition |
|-----------|------------|
| 01. CM    | Abrupt motion of the camera |
| 02. ROT   | Target object rotates in the video |
| 03. DEF   | The target is deformable |
| 04. FOC   | Target is fully occluded |
| 05. LI    | Low illumination |
| 06. OV    | The target completely leaves the video sequence |
| 07. POC   | Partially occluded |
| 08. VC    | Viewpoint change |
| 09. SV    | Scale variation |
| 10. BC    | Background clutter |
| 11. MB    | Motion blur |
| 12. ARC   | The ratio of bounding box aspect ratio is outside the range [0.5, 2] |
| 13. FM    | The motion of the target is larger than the size of its bounding box |
| 14. NMO   | No motion |
| 15. IV    | Illumination variation |
| 16. OE    | Over exposure |
| 17. ROM   | Influence of background object motion for Event camera |

Fig. 5. Distribution of the proposed VisEvent dataset.

diverse ranges, that is, [1–100, 101–300, 301–500, 501–700, 701–1000, 1000+]. We can find that our dataset is suitable for the evaluation of both short-term and long-term tracking. For the challenging factors, we have [136, 28, 8, 73, 75, 64, 114, 57, 131, 211, 33, 70, 90, 33, 63, 38, 239] videos for the 17 attributes listed in Table II, respectively. Our dataset contains many videos with camera motion, background clutter, scale variation, occlusion, and motion of distractors. Experimental results in Section V show that the visual tracking problem in these scenarios is far from being solved.

E. Discussion

In this section, we give a direct comparison between Visible-Event and other dual-modal tracking tasks, including RGB-Thermal and RGB-Depth. These two tasks also attempt to fuse dual modalities for robust object tracking. For the RGB-Thermal, the thermal sensor can sense the temperature of the surface of the object and is not affected by the illumination. Therefore, it has a long sensing distance and works well in the nighttime. However, this sensor is sensitive to thermal cross-over, that is, the image quality is bad when the target object has a similar temperature with background and MB. The high price is also one of the reasons restricting its wide applications. For the RGB-Depth, the depth sensors can perceive objects well in 3-D space, however, it may only work well in local space due to the fact that its sensing distance is limited (usually less than 10 m). Also, it cannot handle the issue of low light and high speed. In contrast, the Event cameras, such as the DVS [10], are bio-inspired vision sensors that output pixel-level brightness changes instead of standard intensity frames. They offer significant advantages over standard cameras, namely, a very high dynamic range, NOM blur, and latency in the order of microseconds. The following tutorials are recommended to have a general understanding of Event cameras.\footnote{https://youtu.be/D6rv6p09XyWU}

The study of object tracking using Event cameras is a new topic, therefore, many problems need to be addressed to achieve reliable object tracking in challenging scenarios. For example, how to represent the event flows to fully exploit the spatiotemporal information, and how to design efficient neural networks like spiking neural networks for effective feature learning. More importantly, there are still no public large-scale realistic visible-event datasets and baseline methods for object tracking which seriously limited the development of this research direction. In this work, we propose a large-scale benchmark dataset termed VisEvent to handle this problem, some sample images are visualized in Fig. 6. In addition, we also construct multiple baseline trackers by extending visible trackers into dual-modality versions.

V. EXPERIMENTS

A. Dataset and Evaluation Metric

In this work, our tracker is trained on the training subset of our proposed VisEvent dataset which contains 500 video sequences. For the testing, we evaluate and compare the trackers on the testing subset (320 videos) of VisEvent, and also two simulated tracking datasets, that is, the OTBDVS and VOT-DVS. We adopt simulation toolkit V2E \footnote{https://zhangjiqing.com/dataset/} to complete the conversion. We also report and compare with other state-of-the-art trackers on FE108,\footnote{https://github.com/Event-AHU/COESOT} and COESOT\footnote{https://github.com/wangxiao5791509/VisEvent_SOT_Benchmark} dataset. The FE108 totally contains 108 videos and the authors separate them into 76 and 32 videos for the training and testing, respectively. The COESOT is a newly released generic RGB-Event tracking dataset that contains 90 categories of target objects and 1354 video sequences. The training and testing subset contains 827 and 527 videos, respectively.

Two popular metrics are adopted for the evaluation of tracking performance, including Precision Plot and Success Plot. Specifically, the Precision Plot illustrates the percentage of frames where the center location error between the object location and ground truth is smaller than a predefined threshold (default value is 20 pixels). Success Plot demonstrates the percentage of frames the IoU of the predicted and the ground truth bounding boxes is higher than a given ratio. The dataset, source code, and evaluation toolkit can be found at https://github.com/wangxiao5791509/VisEvent_SOT_Benchmark.

B. Baseline Construction

To further boost the development of this research area, in this work, we construct many baseline trackers by fusing visible frames and event flows for future works to compare. Representative Trackers: Three kinds of representative tracking frameworks are explored in this work.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
1) **Binary Classification-Based Trackers:** MDNet [19], RT-MDNet [20], VITAL [40], Meta-Tracker [63], and MANet [64].

2) **CF-Based Trackers:** KCF [65], STRCF [66], MOSSE [67], CSK [68], CN [69], DAT [70], and LDES [71].

3) **Siamese Matching-Based Trackers:** SiamFC [72], SiamFC++ [73], SiamRPN [74], SiamRPN++ [53] (AlexNet, ResNet50, and Long-term versions), ATOM [21], DIMP [54], PrDIMP [75], SiamRCNN [76], Ocean [77], and SiamDW [78].

**Various Fusion Strategies:** Three kinds of fusion strategies are considered when extending the aforementioned trackers.

1) **Early Fusion** denotes the strategy that fuses the input data before feeding them into the tracking model. In this article, two kinds of operations are explored, in more detail, we first simply add or concatenate corresponding RGB and event frame as one unified data for tracking. Therefore, existing RGB trackers can be directly tested as baseline algorithms to compare for future works.

2) **Middle Fusion** is also termed Feature Fusion and it is widely used in current multimodal fusion methods. In this work, we consider the following approaches for fusion.

   a) **Concat:** We simply concatenate the features of dual modalities to get the fused representation for tracking.

   b) **Add:** The two features are added together as the final features.

   c) **1 × 1 Conv:** The convolutional layer with kernel 1 × 1 is used for fusing the feature maps.

   d) **CAssen:** The widely used channel attention (CAssen) is employed for fusion.

   e) **SAtt:** SAtt is used for feature fusion.

   f) **CAM:** Cross attention module proposed in [79].

3) **Late Fusion** (or response fusion) targets at combining the response score or activation map output from the tracking model.

**C. Benchmark Comparison**

**Results on VisEvent Dataset:** In this work, we construct multiple baseline methods for future works to compare on our dataset and report part of these results in Fig. 7. Specifically, we can find that the CF-based trackers achieve lower scores on this benchmark due to the manually designed features used in their model, such as HOG, gray pixels, etc. With the help of the deep features, binary classification-based trackers achieve better performance than CF-based trackers. For example, the MDNet [19], VITAL [40], and RT-MDNet [20] get 0.627/0.426, 0.616/0.415, and 0.560/0.352, respectively, while the CN [69] and KCF [65] only achieve 0.458/0.269 and 0.453/0.260.

Interestingly, we also find that the Siamese network-based trackers usually occupy the top few rankings of other datasets, but are not always very prominent on our dataset. Among of them, the SiamRPN++ [53] achieves 0.538/0.379, 0.576/0.410, and 0.539/0.387, when AlexNet [80], ResNet50 [81], and long-term versions are evaluated. These results are significantly worse than MDNet and its multiple extensions which may demonstrate that the number of layers of the backbone network is not the most important.
Fig. 7. Tracking results on the proposed VisEvent dataset (part of the constructed baselines are reported in this figure). EF and MF denote the early and middle fusion strategy used for the extension of the corresponding trackers, respectively.

role for Visible-Event tracking. The long-term version of SiamRPN++ did not increase the overall score when comparing it with its short-term version. This phenomenon fully demonstrates the challenge of our proposed tracking dataset. Compared with these works, we can attain better results with the help of the CMT module, that is, 0.632/0.430 on precision and success plots, respectively. Our results are even better than DiMP50 [54] which is a very strong tracker developed based on deep residual networks (50 layers, Ours: 3 convolutional layers). These results fully demonstrate the effectiveness and advantages of our proposed baseline tracker.

To compare with recent event-based trackers STNet [55] and strong trackers like TransT [23], Ocean [77], in this part, we evaluate our tracker on the subset of VisEvent dataset, termed VisEvent-aedat4 (i.e., the videos with aedat4 file only), by following STNet [55]. As shown in Table III, we can find that our MDNet-based tracker achieves 0.460/0.280 on PR and SR metrics, which is comparable with existing strong trackers, such as TransT [23] and ATOM [21], and better than KYS [82], DiMP [54], etc. Our results are inferior to STNet [55] which is a Swin Transformer-based event tracker proposed by Zhang et al., this may be caused by the fact that the STNet considers the temporal information well using spiking neural networks. In our future works, we will consider designing advanced temporal information mining modules for high-performance tracking.

Results on FE108 Dataset: As shown in Table IV, we compare with multiple strong Siamese tracker on FE108 dataset, including SiamRPN++ [53], SiamBAN [83], SiamFC++ [73], and KYS [82]. We can find that the results of these trackers on this benchmark are poor. Our tracker based on MDNet and ATOM attains 0.578/0.351 and 0.794/0.543, respectively, which are significantly better than these trackers.

| TABLE III |
| EXPERIMENTAL RESULTS (PR|SR) ON THE VISEVENT-AEDAT4 SUBSET |
| STNet [55] | KYS [82] | TransT [23] | SiamRPN [74] |
| STNet [55] | 0.655/0.355 | 0.445/0.313 | 0.471/0.329 | 0.372/0.252 |
| KYS [82] | 0.462/0.291 | 0.404/0.279 | 0.434/0.322 | 0.460/0.280 |

These experiments on this benchmark also validated the advantages of our trackers on RGB-DVS tracking.

Results on COESOT Dataset: As shown in Table V, we report our tracking results and compare them with other strong visual trackers on the recently released COESOT dataset. Obviously, our proposed MDNet-based RGB-Event tracker obtains 0.665/0.533 on the PR and SR metrics, which is significantly better than most of the compared SOTA trackers, such as the SiamBAN-ACM [8], SiamRPN [74], RTS50 [84], and comparable with STARK-ST50 [85] and Mixformer22k [86]. Note that, the latter two are strong Transformer-based trackers proposed in recent years. These comparisons fully demonstrate the effectiveness of our proposed CMT fusion module for RGB-Event visual tracking.

Results on Artificial OTB-DVS [87] and VOT-DVS [88]: To comprehensively validate the effectiveness of our model, we also test it on two popular tracking datasets, including OTB-DVS and VOT-DVS datasets, and report their AUC score in Table VI. Specifically, we can find that our model achieves 0.68 and 0.33 based on RGB videos and Event images, respectively, on the OTB-DVS. Better results can be obtained if both of them are used, that is, 0.69 on this dataset. For the VOT-DVS, we get 0.39, 0.18, and 0.43 on these three settings, respectively. These results fully demonstrate the effectiveness of event flows for the improvement of tracking performance.
It is worth noting that we only conduct self-comparison on the two datasets and do not compare with other trackers, as the two datasets are all simulated data.

D. Ablation Study

Influence of Input Modalities: In this work, we report the tracking results with a single modality to validate the effectiveness of combining visible and event sensors. As shown in Table VII, the baseline tracker MDNet achieves 0.605/0.412 when only visible frames are used. If only the event frames are used, it can achieve 0.460/0.280. It is significantly worse than tracking results with visible cameras which demonstrates that only the event sensors are not enough for practical tracking. Because it can only capture where events occurred in the scene and provide outline shape information. This information is very important for tracking but we still need the appearance and detailed texture information to discriminate the target object from other distractors. When we combine both modalities for tracking, the overall performance can be improved to 0.627/0.426.

In addition, we also test the PrDIMP18 tracker based on RGB and event images only and get the 0.554/0.407 and 0.404/0.256, respectively. When we fuse the two modalities with early fusion, we can get 0.578/0.421, which is significantly better than a single modality only. It also attained better results on the MB and LI attributes. These experimental results fully demonstrate the useful clues provided by event flows. Similar conclusions can also be drawn from the results based on RT-MDNet. Therefore, it will be an interesting research direction of reliable object tracking through the collaboration of video frames and event flows.

Influence of CMT: To better understand the contributions of the feature fusion module in our proposed trackers, we integrate it with MDNet and RT-MDNet to check its influence on final tracking. The tracking results are reported in Table VII. When integrated into MDNet, clearly, its results 0.627/0.426 can be improved to 0.632/0.430. It also helps RT-MDNet by improving the results from 0.560/0.352 to 0.574/0.359. These results demonstrate the effectiveness of the feature-level information fusion module CMT for tracking.

Influence of Challenging Factors: In this work, 17 attributes are considered for the VisEvent tracking dataset. In this section, we report most of them in Fig. 8, including LI and FM. It is easy to find that our proposed tracker attains the top-5 even the best tracking performance under these attributes. These results demonstrate the effectiveness of our proposed modules for tracking under extremely challenging scenarios. More results can be found on our project page.

E. Results of Various Fusion Strategies

In this section, we test different feature fusion strategies based on MDNet [19] for RGB-Event visual tracking, as shown in Table VIII. To be specific, the widely used fusion methods, such as concatenate, add, and convolution fusion

| Method | concat | add | 1 x 1 Conv | CRAM | SAT | CAM | CRAM (Ours) |
|--------|--------|-----|------------|------|-----|-----|------------|
| Prt. Fis. | 0.627  | 0.471 | 0.617  | 0.456 | 0.466 | 0.696 | 0.632 |
| Int. Fis. | 0.626  | 0.474 | 0.624  | 0.477 | 0.470 | 0.692 | 0.430 |
Fig. 9. Visualization of tracking results of our proposed tracker and other SOTA trackers on the VisEvent dataset.

with kernel size $1 \times 1$. The $C_{Att}$, $S_{Att}$, cross-attention (CAM) used in the work [79], are also exploited for multimodal fusion. According to Table VIII, we can find that the simple concatenate of dual features can bring the best tracking performance, that is, 0.627/0.426. Interestingly, the add, $C_{Att}$, and $S_{Att}$ all achieve inferior results. We think this may be caused by the fact that the event images only provide shape information and it may hurt the visible features by simple add, and two modal information are saved to the greatest extent with the concatenating operation.

F. Efficiency Analysis

The proposed CMT is a general module for visible-event tracking, therefore, its efficiency mainly depends on the used baseline tracker. For example, when integrating our CMT into dual-modality RT-MDNet, it can run at about 14 FPS. With the help of CMT, we achieve the best tracking performance on the proposed benchmark dataset.

G. Visualization

In addition to the aforementioned quantitative analysis, in Fig. 9, we also give some visualization of our tracking results and compared trackers. We can find that the current strong tracker PrDiMP50 [75], SuperDiMP [54], and SiamRPN++ [53] still suffer from MB, FM, LI, etc. The online trackers MDNet [19], VITAL [40], and ours can handle these scenarios with the help of event images.

H. Failed Case Analysis

Although this article achieves good results on some videos of our dataset, however, this problem is still far from being solved. The two failed cases we provided in Fig. 10 are baseball and moving stars. We can find that the baseball is a fast-moving object and is almost invisible in the RGB camera. The event camera captures the moving baseball well and clearly shows the trajectory compared with the RGB frames. However, the event image representation used in this article may be a suboptimal choice which may lead to shape variation, as shown in the second row. For the moving stars, similar issues occurred in this case, the stacked event streams overlapped between different stars. We believe the event image representation stacked in a fixed time window is the key reason. In our future works, we will consider adopting other event representations like event points or voxels to better capture the spatiotemporal information for visual object tracking.

VI. CONCLUSION AND FUTURE WORKS

In this article, we propose a new and large-scale object tracking benchmark by combining the visible and event sensors. It targets providing the characteristic of high dynamic range and high temporal resolution for standard visible cameras with biologically inspired event cameras. This will widely extend the applications of current visual trackers in practical scenarios, such as high speed, low light, and cluttered backgrounds. Our dataset contains 820 video pairs that are collected with DVS cameras and it involves multiple types of objects and scenarios. We provide multiple baseline methods by extending current state-of-the-art trackers into dual-modality versions. In addition, we also designed a simple but effective baseline tracker by developing CMT modules for interactive feature learning and fusion. Extensive experiments on the proposed VisEvent dataset fully demonstrate its good performance. We hope this work will boost the development of object tracking based on neuromorphic cameras. In our future works, we will continue to explore new architectures of pure spiking neural networks for this tracking task.

ACKNOWLEDGMENT

The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

REFERENCES

[1] Z. Zhang, Y. Liu, X. Wang, B. Li, and W. Hu, “Learn to match: Automatic matching network design for visual tracking,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 13339–13348.
[2] X. Wang et al., “Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13763–13773.
[3] X. Wang, J. Tang, B. Luo, Y. Wang, Y. Tian, and F. Wu, “Tracking by joint local and global search: A target-aware attention-based approach,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 11, pp. 6931–6945, Nov. 2022.
[4] X. Wang et al., “Dynamic attention guided multi-trajectory analysis for single object tracking,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 12, pp. 4895–4908, May 2021.
X. Dong, J. Shen, L. Shao, and F. Porikli, “CLNET: A compact latent network for fast adjusting Siamese trackers,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 378–395.

J. Shen, Y. Liu, X. Dong, X. Lu, F. S. Khan, and S. Hoi, “Distilled Siamese networks for visual tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8896–8909, Dec. 2021.

Z. Zhou, H. Liu, D. P. Moeys, G. Das, D. Neil, S.-C. Liu, and T. Delbrück, “Event-driven stereo visual tracking algorithm to solve object occlusion,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 4, pp. 1053–1054, 2022.

B. Ramesh et al., “E-TLD: Event-based framework for dynamic object tracking,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3996–4006, May 2020.

L. A. Camuñas-Mesa, T. Serrano-Gotarredona, S.-H. Ieng, R. Benosman, and B. Linares-Barranco, “Event-driven stereo visual tracking algorithm to solve object occlusion,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 20, pp. 10534–10541, 2020.

X. Li, Q. Liu, N. Fan, Z. Zhou, Z. He, and X.-Y. Jing, “Dual-regression model for visual tracking,” Neural Netw., vol. 132, pp. 364–374, Dec. 2020.

X. Li, Q. Liu, N. Fan, Z. He, and H. Wang, “Hierarchical spatial-aware Siamese network for thermal infrared object tracking,” Knowl. Based Syst., vol. 166, pp. 71–81, Feb. 2019.

X. Dong, J. Shen, L. Shao, H. Ling, and F. Porikli, “Dynamical hyperparameter optimization via deep reinforcement learning in tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 5, pp. 1515–1529, May 2021.

X. Lu, C. Ma, J. Shen, X. Yang, I. Reid, and M.-H. Yang, “Deep object tracking with shrinkage loss,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, pp. 2380–2396, May 2022.

Y. Han, C. Deng, Y. Fu, B. Yin, and B. Dong, “Object tracking by jointly exploiting frame and event domain,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 13043–13052.

H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4293–4302.

I. Jung, J. Son, M. Baek, and B. Han, “Real-time MDNET,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 83–91.

M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ATOM: Accurate tracking by overlap maximization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4660–4669.

Y. Chen, B. Liu, L. Liu, and J. Tang, “SINT++: Robust visual tracking via adversarial positive instance generation,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13043–13052.

X. Wang, C. Li, B. Luo, and J. Tang, “SINT++: Robust visual tracking via adversarial positive instance generation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 8890–8899.

Q. Guo, W. Feng, Z. Chen, R. Gao, L. Wan, and S. Wang, “Effects of blur and deblurring to visual object tracking,” 2019, arXiv:1908.07904.

B. Ramesh, S. Zhang, W. Lee, Z. Gao, G. Orchard, and C. Xiang, “Long-term object tracking with a moving event camera,” in Proc. IEEE Trans. Cybern., vol. 53, no. 3, pp. 1968–1981, Mar. 2023.

W. O. C. Hernandez, J. Andrade-Cetto, and J. S. Ortega, “High-speed event camera tracking,” in Proc. 31st Brit. Mach. Vis. Virtual Conf., 2020, p. 1–12.

I. A. Lopez and M. Chli, “HASTE: Multi-hypothesis asynchronous speeded-up tracking of events,” in Proc. 31st Brit. Mach. Vis. Virtual Conf., 2020.

Z. Cao, L. Cheng, C. Zhou, N. Gu, X. Wang, and M. Tan, “Spiking neural network-based target tracking control for autonomous mobile robots,” Neurocomput. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP), 2017, pp. 6000–6010.

B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “SIAMRPN++: Evolution of Siamese visual tracking with very deep networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4282–4291.

G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, “Learning discriminative model prediction for tracking,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6182–6191.

J. Zhang et al., “Spiking transformers for event-based single object tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 8801–8810.
WANG et al.: VisEvent: RELIABLE OBJECT TRACKING VIA COLLABORATION OF FRAME AND EVENT FLOWS 2009

...networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

[58] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck, “DVS benchmark datasets for object tracking, action recognition, and object recognition,” Front. Neurorosci., vol. 10, p. 405, Aug. 2016.

[59] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, “Event-based moving object detection and tracking,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2018, pp. 1–9.

[60] L. Huang, X. Zhao, and K. Huang, “GOT-10K: A large high-diversity benchmark for generic object tracking in the wild,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 5, pp. 1562–1577, May 2021.

[61] 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., 2021, pp. 1312–1321.

[62] E. Park and A. C. Berg, “Meta-tracker: Fast and robust online adaptation for visual object tracks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 587–604.

[63] C. Li, A. Lu, A. Zheng, Z. Tu, and J. Tang, “Multi-adapter RGBT tracking,” in Proc. ICCV Workshops, 2019, pp. 2262–2270.

[64] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High-speed tracking with kernelized correlation filters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.

[65] F. Li, C. Tian, W. Zao, L. Zhang, and M.-H. Yang, “Learning spatial-temporal regularized correlation filters for visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4904–4913.

[66] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2010, pp. 2544–2550.

[67] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 702–715.

[68] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van de Weijer, “Adaptive color attributes for real-time visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1090–1097.

[69] H. Possegger, T. Mauthner, and H. Bischof, “In defense of color-based model-free tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 2113–2120.

[70] Y. Li, J. Zhu, S. C. Hoi, W. Song, Z. Wang, and H. Liu, “Robust estimation of similarity transformation for visual object tracking,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8666–8673.

[71] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, “Fully-convolutional Siamese networks for object tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 850–865.

[72] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, “SIAMFC++: Towards robust and accurate visual tracking with target estimation guidelines,” in Proc. AAAI, 2020, pp. 12549–12556.

[73] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with Siamese region proposal network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8971–8980.

[74] M. Danelljan, L. V. Gool, and R. Timofte, “Probabilistic regression for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 7183–7192.

[75] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, “SIAM R-CNN: Visual tracking by re-detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6578–6588.

[76] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, “OCEAN: Object-aware anchor-free tracking,” in Proc. 16th Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, pp. 771–787.

[77] Z. Zhang and H. Peng, “Deeper and wider Siamese networks for real-time visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4591–4600.

[78] W. Sue, M. Sun, P. Wang, and Q. Wu, “Proposal-free one-stage referring expression via grid-word cross-attention,” in Proc. IJCAI, 2021, pp. 1–8.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 25, 2012, pp. 1097–1105.

[80] K. S. X. Xiang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[81] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, “Know your surroundings: Exploiting scene information for object tracking,” in Proc. 16th Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, pp. 32–41.

[82] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, “Siamese box adaptive network for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6668–6677.

[83] M. Paul, M. Danelljan, C. Mayer, and L. Van Gool, “Robust visual tracking by segmentation,” in Proc. 17th Eur. Conf. Comput. Vis. (ECCV), Oct. 2022, pp. 571–588.

[84] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, “Learning spatio-temporal transformer for visual tracking,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10448–10457.

[85] Y. Cui, C. Jiang, L. Wang, and G. Wu, “MixFormer: End-to-end tracking with iterative mixed attention,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 13608–13618.

[86] Y. Wu, J. Lim, and M.-H. Yang, “Object tracking benchmark,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[87] Y. Kaspi et al., “The seventh visual object tracking VOT2019 challenge results,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, p. 10.

Xiao Wang (Member, IEEE) received the B.S. degree from the College of Computer and Information Technology, China Three Gorges University, Yichang, China, in 2014, the M.S. degree from the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China, in 2017, and the Ph.D. degree from the National Engineering Research Center for Visual Technology, School of Computer Science, Peking University, Beijing, China, in 2022.

He has authored or coauthored over 20 technical papers in refereed journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, CVPR, ICCV, AAAI, ECCV, ACCV, ACM-MM, and WACV. For more information, see: https://xwanga5791509.github.io/.

Lin Zhu (Member, IEEE) received the B.S. degree in computer science from Northwestern Polytechnical University, Xi’an, China, in 2014, the M.S. degree in computer science from North Automatic Control Technology Institute, Taiyuan, China, in 2018, and the Ph.D. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2022. He is currently an Assistant Professor with the School of Computer Science, Beijing Institute of Technology, Beijing. His current research interests include image processing, neuromorphic computing, and spiking neural network.
Zhipeng Zhang received the bachelor’s degree in optics and information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2022. He currently works as a Senior Autodriving Researcher with DiDi. His current research interest is video and perception understanding, including object tracking, object segmentation, video representation learning, monocular, and multiview 3-D perception.

Zhe Chen (Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2014, and the Ph.D. degree from the UBTech Sydney Artificial Intelligence Centre, the Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia, in 2019. His studies were published in IEEE CVPR, ICIAP, ECCV, and Journal of the Atmospheric Sciences. His research interests include object detection, computer vision, and deep learning.

Xin Li received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2012, 2014, and 2020, respectively. He joined Peng Cheng Laboratory, Shenzhen, China, in September 2022. His current major research interests are computer vision and deep learning.

Yaowei Wang (Member, IEEE) received the Ph.D. degree in computer science from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2005. He is currently an Associate Professor with Peng Cheng Laboratory, Shenzhen, China, and a Professor with the National Engineering Laboratory for Video Technology Shenzhen, Peking University, Beijing, China, in 2002. He has authored or coauthored over 70 refereed journals and conference papers. His research interests include machine learning, multimedia content analysis, and understanding.

Dr. Chen serves as a Reviewer for a number of journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON CYBERNETICS.

Yonghong Tian (Fellow, IEEE) received the master’s degree from the University of Electronic Science and Technology of China, Chengdu, China, in June 2000, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in July 2005. He is currently a Boya Distinguished Professor with the Department of Computer Science and Technology, Peking University, Beijing, and is also the Deputy Director of Artificial Intelligence Research Center, Peng Cheng Laboratory, Shenzhen, China. He has authored or coauthored over 200 technical articles in refereed journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE/IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS/IEEE TRANSACTIONS ON IMAGE PROCESSING/IEEE TRANSACTIONS ON MULTIMEDIA/IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY/IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING/IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, ACM Computing Survey/ACM Transactions on Information Systems/ACM Transactions on Multimedia Computing, Communications, and Applications and conferences, such as NeurIPS/CVPR/ICCV/AAAI/ACMMM/WWW. His research interests include neuromorphic vision, brain-inspired computation, and multimedia big data.

Feng Wu (Fellow, IEEE) received the B.S. degree in electrical engineering from Xidian University, Xi’an, China, in 1992, and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1996 and 1999, respectively. He is currently a Professor with the University of Science and Technology of China, Hefei, China. Before that, he was a Principle Researcher and a Research Manager with Microsoft Research Asia, Beijing, China. He has authored or coauthored over 200 high-quality papers (including several dozens of IEEE Transaction papers) and top conference papers on MOBICOM, SIGIR, CVPR, and ACM MM. He has 77 granted U.S. patents. His 15 techniques have been adopted into international video coding standards. His research interests include image and video compression, media communication, and media analysis and synthesis.

Prof. Wu got the Best Paper Award in IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY 2009, the Progress in Construction Materials 2008, and SPIE VCIP 2007. He got the IEEE Circuits and Systems Society 2012 Best Associate Editor Award. He serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON MULTIMEDIA. He also serves as the TPC Chair for MMSPI 2011, VCIP 2010, and PCM 2009, and the Special Sessions Chair for ICME 2010 and ISCAS 2013.