Event-Assisted Object Tracking on High-Speed Drones under Harsh Illumination Environment

Yuqi Han , Xiaohang Yu , Heng Luan , Jinli Suo *

Posted Date: 14 December 2023

doi: 10.20944/preprints202312.1056.v1

Keywords: Drones; harsh illumination; image enhancement; event-assisted object tracking; multi-sensor fusion

Article

Event-Assisted Object Tracking on High-Speed Drones under Harsh Illumination Environment

Yuqi Han [1], Xiaohang Yu [2], Heng Luan [3] and Jinli Suo [1]

[1] Department of Automation, Tsinghua University
[2] Tsinghua-UC Berkeley Shenzhen Institute
[3] Research and Development Center, TravelSky Technology Limited
* Correspondence: jlsuo@tsinghua.edu.cn (J.S.)

Abstract: Drones have been used in a variety of scenarios such as atmospheric monitoring, fire rescue, agricultural irrigation, etc., in which accurate environmental perception is of crucial importance for both decision-making and control. Among the drone sensors, the RGB camera is indispensable for capturing rich visual information for vehicle navigation but encounters a grand challenge in high-dynamic-range scenes that occur frequently in real applications. Specifically, the recorded frames suffer from under-exposure and over-exposure simultaneously and degenerate the successive vision tasks. To solve the problem, we take object tracking as an example and leverage the superior response of event cameras over a large intensity range to propose an event-assisted object tracking algorithm that can achieve reliable tracking under large intensity variations. Specifically, we propose to pursue feature matching from dense event signals, and based on which to (i) design a UNet-based image enhancement algorithm to balance the RGB intensity with the help of neighboring frames in the time domain, and then (ii) construct a dual-input tracking model to track the moving objects from intensity balanced RGB video and event sequence. The proposed approach is comprehensively validated in both simulation and real experiments.

Keywords: drones; harsh illumination; image enhancement; event-assisted object tracking; multi-sensor fusion

1. Introduction

As a lightweight, flexible, and cost-effective [1–3] platform, drones have often been used in a variety of remote tasks, such as surveillance [4,5], detection [6], and delivery [7]. In such applications, drones need to accurately perceive the surrounding environments to support subsequent decisions and actions. In general, the common sensors used on UAVs include visible-wavelength optical cameras [8], LiDAR [9], NIR/MIR cameras [10], etc. Each type of sensor has its own advantages and disadvantages, so multi-mode sensing has been the typical solution in this field. Among the various sensors, visible-wavelength camera is an indispensable sensing unit due to its high resolution, capability of collecting rich information, and low cost of construction.

As one of the most important tasks of a drone, object tracking [11–14] has been widely studied. Broadly speaking, the object-tracking algorithms take either the RGB frame as input or its combination with other sensing modes. The RGB-only methods [15–18] prevail in frame-based object tracking, but are limited in hash-illumination scenarios. Some researchers proposed to incorporate information from event-based cameras, which show superior performance in both low-light and high-dynamic-range scenes. To fuse the information from RGB frames and event sequence, Mitrokin et al. [19] proposed a time-image representation to combine temporal information of the event stream, and Chen et al. [20] improved the event representation by proposing a synchronous Time-Surface with Linear Time Decay representation. These approaches exhibit promising performance in object tracking with high time consistency. However, the above methods are difficult to apply on 24/7 UAVs due to the limited sensing capability of RGB sensors in cases with complex illumination. Especially, for scenarios with high dynamic range lighting, overexposure and underexposure coexist and the image quality degrades

 doi:10.20944/preprints202312.1056.v1

greatly, which hampers the successive analysis. For object tracking, it is difficult to capture consistent image features in the time domain, and therefore performance is degraded. Taking the video in Figure 1 as an example, when capturing a car traveling through a tunnel there existing large intensity range in each frame and abrupt variation in different frames. In the latter frames, the car is even undetectable in some frames due to underexposure, for both tracking algorithms and human vision systems.

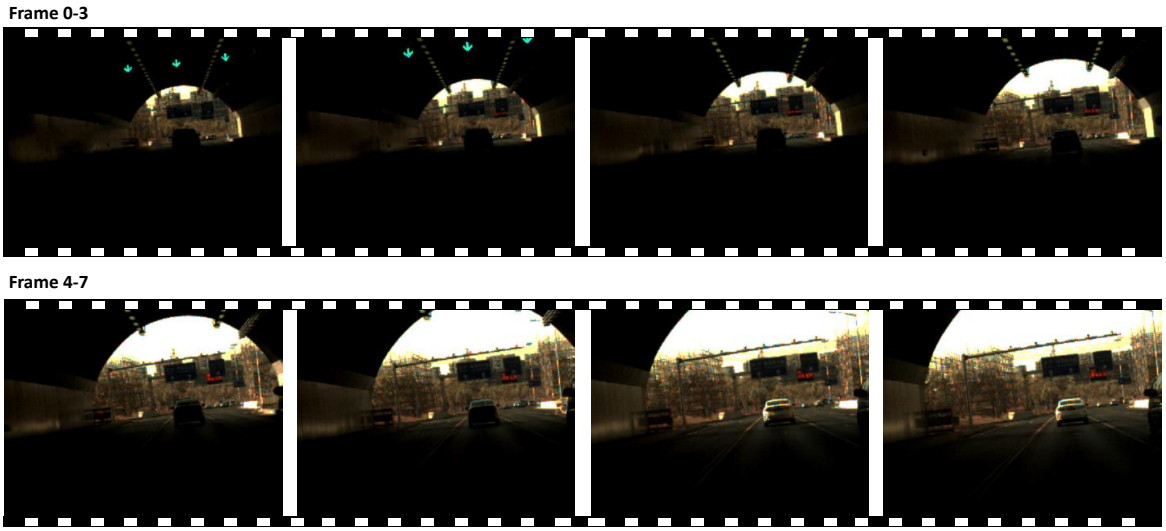

Figure 1. A typical high-dynamic-range RGB video of a car driving through the tunnel, in which the car is almost invisible in the last 5∼6 frames due to underexposure.

To raise the image quality under harsh illumination, researchers have made a lot of explorations in recent years. One most common way is to reconstruct HDR images by merging the set of multi-exposure LDR images [21]. For dynamic scenes, image alignment is required to address the inconsistency between frames with different exposures. Kang et al. [22] initially aligned neighboring frames with the reference frame, and merged these aligned images to craft an HDR image. Later works [23,24] modified it by adding a motion estimation block and a refinement stage. Differently, Kalantari et al. [25] proposed a patch-based optimization technique, synthesizing absent exposures within each image before reconstructing the ultimate HDR image. Gryaditskaya et al.[26] enhanced this method by introducing an adaptive metering algorithm capable of adjusting exposures, thereby mitigating artifacts induced by motion. Instead of capturing frames with different exposure times, some methods use deep neural networks to reconstruct the HDR image from a single input image. However, relying on a fixed reference exposure, the reconstruction is strongly ill-posed and can not achieve high between-frame consistency. Besides, many existing HDR video reconstruction methods focus on developing some special hardware, such as scanline exposure/ISO [27–29], per-pixel exposure [30] and modulo camera [31], etc., but these new cameras are still in research and not ready for commercial use in a near future. Some other recent approaches work under the deep-optics scheme and focus on jointly optimizing both the optical encoder and CNN-based decoder for HDR imaging challenges. The above methods usually make assumptions about the lighting conditions, which might not hold in real scenes. Besides, most of these algorithms need ground-truth high dynamic range images for supervised network training and of limited performance in scenes different from the training data. Hence, these methods are enlightening but difficult to directly apply on practical UAV platforms working in open environments.

Event camera, also known as neuromorphic vision sensor, is an emerging technique that records intensity changes exceeding the threshold asynchronously [32,33]. In recent years, event signals have been used in a variety of high-speed tasks due to their high sensitivity and fast response, such as high-speed tracking [34–37], frame interpolation [38,39], optical flow estimation [40–42], motion

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

detection [43], etc. Unlike conventional optical camera sensors, event cameras output the "events" indicating that there occur sufficiently large intensity variation at certain positions and instants, as well as the polarity of the change. Considering that the event camera can record the motion over a large intensity range and is insensitive to abrupt intensity changes, we propose to use the event signals to explicitly align the RGB frames and thus compensate for the quality degradation harming the successive object tracking. Specifically, from the event signals that are robust to abrupt intensity changes, we match the key points occurring at different instants [44] and utilize the matching to balance the intensity change in sequential RGB frames. Afterward, we construct a fusion network to aggregate the enhanced RGB frames and event signals to accomplish robust object tracking.

The contributions of this paper are as follows:

- We propose an event-assisted robust object tracking algorithm working in high-dynamic-range scenes, which successfully integrates the information from an event camera and an RGB camera to overcome the negative impact of harsh illumination on the tracking performance.
- We construct an end-to-end deep neural network to enhance the high-dynamic-range RGB frames and conduct object tracking sequentially, and the model is built in an unsupervised manner.
- We design an approach to match the feature points occurring at different time instants from the dense event sequence, which guides the intensity compensation in high-dynamic-range RGB frames.
- The approach demonstrates superb performance in a variety of harshly lit environments, which validates the effectiveness of the proposed approach and largely broadens the practical applications of drones.

2. Framework and Algorithm Design

This section presents the details of the proposed event-assisted robust object tracking approach working under harshly lit illuminations. Here we first briefly introduce the framework and then describe the design of three key modules, including the retrieval of feature registration across frames, the enhancement of high-dynamic-range frames, and the successive dual-modal object tracking.

The basic idea of the proposed approach is to utilize the reliable perception capability of motion cues by event cameras to resist the quality degradation of RGB frames, and then combine the event signals and the enhanced RGB video to boost the successive tracking performance suffering from overexposure and underexposure. The whole framework of the proposed event-assisted object tracking approach is shown in Figure 2, which consists of mainly three key modules.

i) Retrieving the motion trajectories of key feature points from the dense event sequence. We divide the event sequence into groups occurring in overlapping short time windows, and the key points from Harris corner detection in each event group can construct some motion trajectories. Further, we integrate these short local trajectories to figure out the motion over a longer period across the RGB frames, even under harsh illumination.

ii) Enhancing the high-dynamic-range RGB frame according to the inter-frame matching and information propagation. Based on the matching among feature points across frames, we build a deep neural network to compensate for the overexposure or underexposure regions from neighboring frames with higher visibility reference frames to guide low-visibility objective frames. In implementation, we build a U-Net-based neural network for image enhancement.

iii). Tracking the target objects by fusing information from both RGB and event inputs. We design a tracking model taking dual-modal inputs to aggregate the information from the enhanced RGB frames and event sequence to locate the motion trajectories. Specifically, we construct 3D CNNs for feature extraction, fuse the features from two arms using the self-attention mechanism, and then employ an MLP to infer the final object motion.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

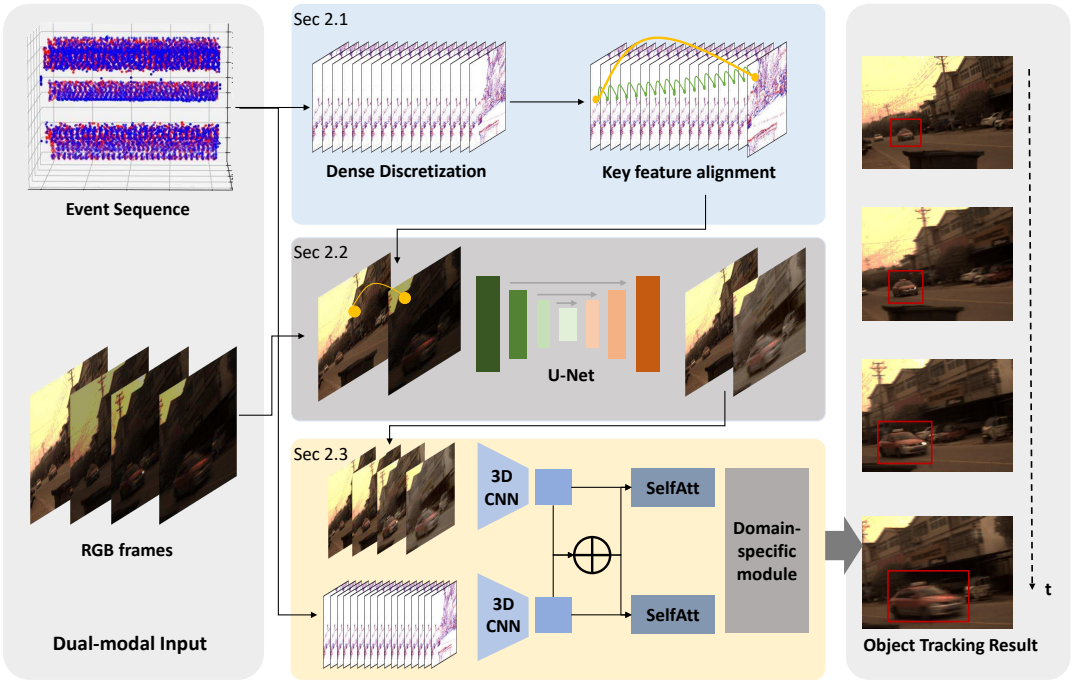

Figure 2. The framework and working flow of the event-assisted robust tracking under harsh illumination. The whole pipeline is fully automatic and consists of three key steps, with the first one by conventional optimization and the latter two implemented by deep neural networks.

2.1. Event-based cross-frame alignment

Event-based key feature extraction and matching are conducted here to utilize the stable event signals under harsh illumination for cross-frame alignment of the degraded RGB video, facilitating the frame compensation from corresponding positions with decent quality in neighboring frames. We locate the key features of moving objects from the event sequence, as illustrated in Figure 3.

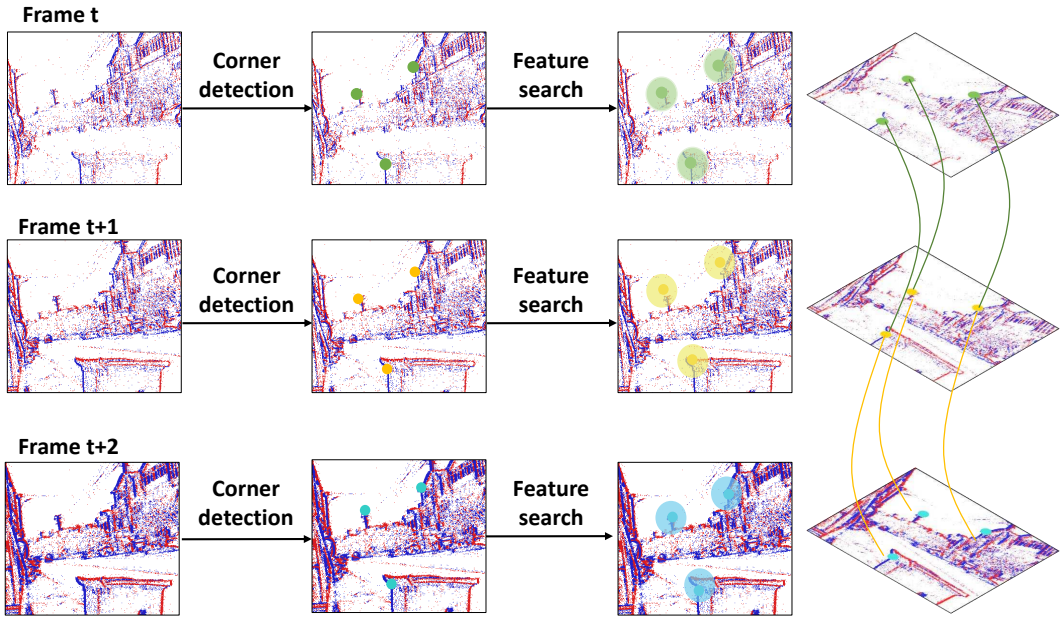

Figure 3. The illustration of the event-based key point alignment. We locate the key feature points via Harris detection and search the matching counterparts locally (circular candidate regions are highlighted with different colors) to constitute the motion trajectories, as shown in the right column.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

Given a time duration T, we assume that there are N RGB frames and S event signals. We define the captured RGB frames as $\{I_0, I_1, \cdots, I_N\}$, and the corresponding time stamps are defined as $\{T_0, T_1, \cdots, T_N\}$. The event signal s is defined as a quadruple x_s, y_s, t_s, P_s, where x_s, y_s denotes the coordinate of s, t_s presents the response time instant, and P_s indicates the polarity of intensity change. Firstly, we divide the S event signals into $K \times N$ groups along the time dimension and project each group into $K \times N$ 2D images, named an event frame. We adopt the Harris corner detection algorithm for the above event frames to extract individual key feature points. Further, we align the key feature points at different time instants. Assuming that the shape of the moving objects is fixed within a very short time slot, i.e., the key features in adjacent frames are similar, we construct a small circular search region with radius r around each key feature. In other words, the key feature at eth frame matches the features inside the searching circle of $e+1$th frame.

For the nth RGB frame, we first align the event frames between $n \times S$ and $(n+1) \times S$. From the displacement between the features of multiple event frames, one can construct the moving trajectory of the key event feature points, which reflects the displacement of the corresponding key features in the RGB frame. Naturally, we can eventually infer the position of the corresponding key feature from nth to $n+1$th RGB frames.

2.2. RGB image enhancement

After matching the feature points in different RGB frames, we enhance the underexposure and overexposure regions utilizing the high-visibility counterparts to adjust the intensity and supplement the details. For description simplicity, we define the low-visibility frames as the objective and the high-visibility frames as the reference. To achieve enhancement, there are two core issues to be addressed: i) how to determine the objective frame that needs to be enhanced; ii) how to design the learning model to improve the visibility to match the reference frame while preserving the original structure of the objective frame.

We first estimate the visibility of the frames to determine which frames are highly degraded. Intuitively, since harsh illumination leads to local overexposure or underexposure which are usually of lacking texture, we use the information richness to characterize the degeneration degree. In implementation we define the visibility V_i of input RGB image R_i as the difference from its low-pass filtered version $\hat{R}_i$, i.e.,

$$V_i = Var(R_i - \hat{R}_i), \tag{1}$$

where $Var(\cdot)$ denotes the variance calculation.

In general, we divide the frames into groups and conduct compensation within each group. We iteratively find the objective frame with the lowest visibility score and the reference frame with the highest visibility, and then conduct enhancement. The iteration ends when the number of iterations exceeds a predetermined number P or the difference between the visibility of the target and the reference frame smaller than η. In our experiments, we set $P = 10$ and $\eta = 0.1$.

For enhancement, we designed a U-Net-shaped network structure inferring the enhanced frame from the objective and reference frames, as shown in Figure 4. The network consists of a three-layer encoder for feature extraction and a three-layer decoder. Skip connections are used to facilitate the preservation of spatial information. The network is trained in an unsupervised manner. We define the loss function based on aligned feature points. Considering that the enhanced frame is expected to be similar to the reference image around the key feature points, and close to the original frame at other locations, we define a combinational loss function. To guarantee the former similarity, we minimize the MSE difference and for the latter we use the LPIPS loss. Denoting the reference image as I_{ref}, the original objective image as I_{obj}, and the output as I_{out}, we define the loss function as

$$L = MSE(I_{\text{ref}}(k) - I_{\text{out}}(k)) + \alpha LPIPS(I_{\text{obj}}(\neg k) - I_{\text{out}}(\neg k)) \tag{2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

where k denotes the positions of key features and $\neg k$ denotes the rest pixels; α is the hyper-parameter that is set to be 0.05 during training.

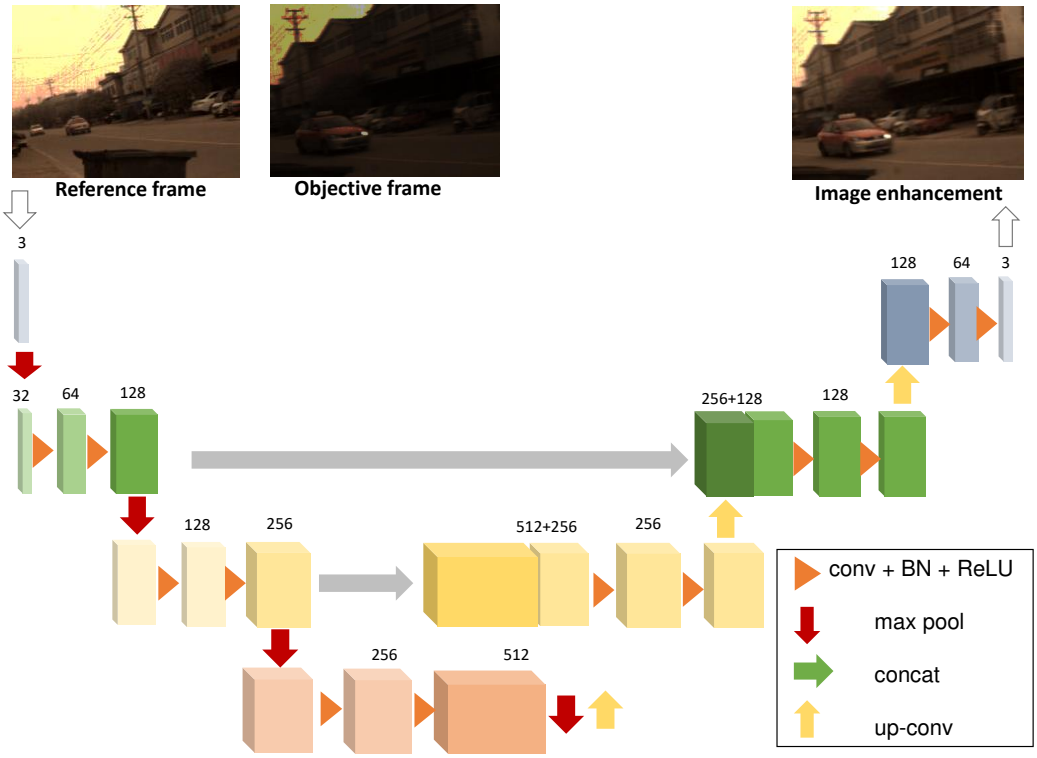

Figure 4. The structure of the RGB image enhancement module. We input the captured RGB frame into the U-Net network, which comprises a three-layer encoder for feature extraction and a three-layer decoder for image enhancement. The network includes skip connections to connect encoder and decoder layers, facilitating the preservation of spatial information. This diagram showcases convolutional, pooling, upsampling, and downsampling layers.

2.3. Dual-modal object tracking

To leverage the motion cues in both the event sequence and the enhanced RGB frames, we construct a dual-modal tracking module for reliable object tracking. The proposed dual-modal tracking module is based on RT-MDNet[45]. The module consists of a shared feature mapping network aiming at constructing the shared representation to distinguish the object from the background, and a domain-specific network focusing on domain-independent information extraction. Different from RT-MDNet[45], the proposed dual-modal design focuses on the feature fusion from two types of inputs and constructs two self-attention modules to highlight the combinational representation from two individual inputs.

The architecture of the network is shown in Figure 5. We first construct two individual 3D CNNs to extract features from the inputs and output feature vectors of the same size. Subsequently, we concatenate the two feature vectors and use convolution to obtain a combinational representation of the fused features. Subsequently, we construct the self-attention network to retrieve the information underlying independent feature inputs. A two-layer fully connected MLP is used to output the common feature. We refer to RT-MDNet[45] to construct the domain-specific layer afterward, outputting the final tracking results.

During model training, for each detection bounding box, a cross-entropy loss function is constructed to ensure that the target and background are separated as much as possible, and multiple domains as well. In the latter fine-tuning stage, we apply different strategies for the first frame and the subsequent ones of a given sequence. For the first frame, we choose multiple bounding boxes following

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

Gaussian distribution to conduct domain-specific adaption, while for the subsequent frames, we build random samples based on the results in the previous frame and search for the proper bounding box by regression.

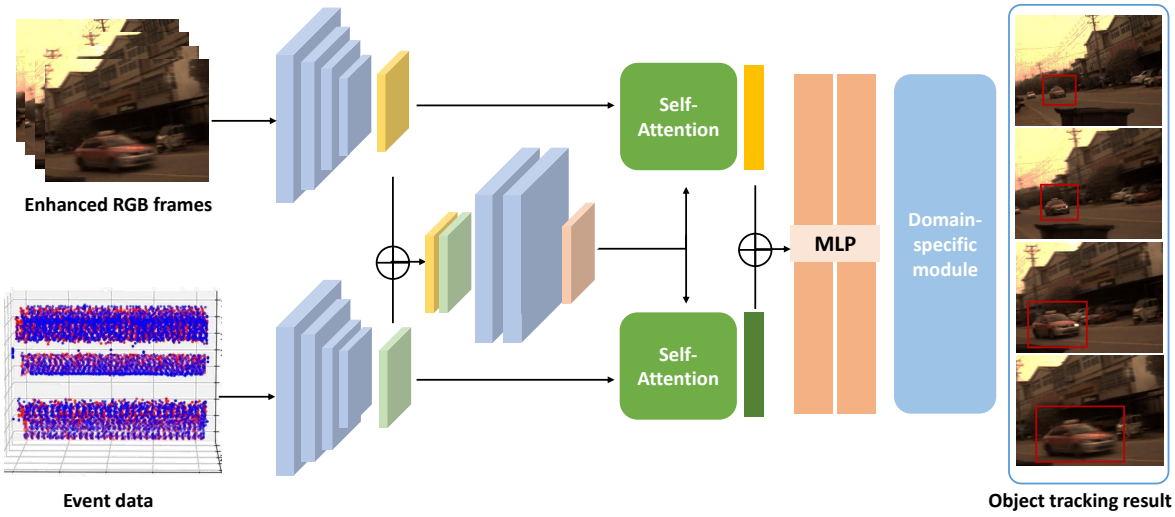

Figure 5. The structure of the object tracking module. The RGB frames and event sequence are individually fed into two 3D CNN modules for feature extraction, and the extracted features are concatenated and sent to another CNN module for fusion. Then, the individual and fused features are separately sent to the self-attention network. Finally, two MLPs are applied to derive the object detection and tracking results.

3. Results

In this section, we construct a series of experiments to verify the effectiveness of the proposed method on two tasks in high-dynamic-range scenes: image enhancement and object tracking. We first present the training details and the datasets. Then we show the visual and quantitative performance against some baseline algorithms. Finally, we conduct ablation experiments to quantify the contribution of the key module of the algorithms.

3.1. Experiment settings

Datasets. We verify the proposed method on both simulated and real datasets. We use the VisEvent[46] as the simulated data and mimic the harsh illumination by modifying the brightness and contrast of the RGB frames. Specifically, we modify the luminance and contrast as follows: we let the luminance vary linearly, quadratically, or exponentially across the frames, and the image contrast undergoes a linear change with different slopes. We first randomly select 1/3 of the data for luminance modification and then apply contrast modification to 1/3 randomly selected videos. Two examples from the simulated dataset are shown in Figure 6. The first scene mimics the brightness changes of the underexposed scenes and the second scene simulates the overexposure, by modifying the image brightness and contrast. One can see that we can generate videos under complex illuminations from the original counterpart with roughly uniform illuminance. In the generated high-dynamic-range RGB frames, the textures of some regions are invisible in some frames due to either underexposure or overexposure. In contrast, the contours across the whole field of view are recorded decently.

For the real-world dataset, we capture some typical nighttime traffic scenes with a pair of registered cameras (one RGB and the other events). The scenes consist of complex illuminations (e.g., traffic lights, neon signs, etc.) and large intensity variations. From the two exemplar scenes in Figure 7, these scenarios exhibit large illuminance variations and the traffic participants are almost invisible in some frames, due to either underexposure or overexposure. This challenging dataset can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

be directly used to test the effectiveness of the proposed object tracking algorithm in real scenarios, as shown in Figure 7.

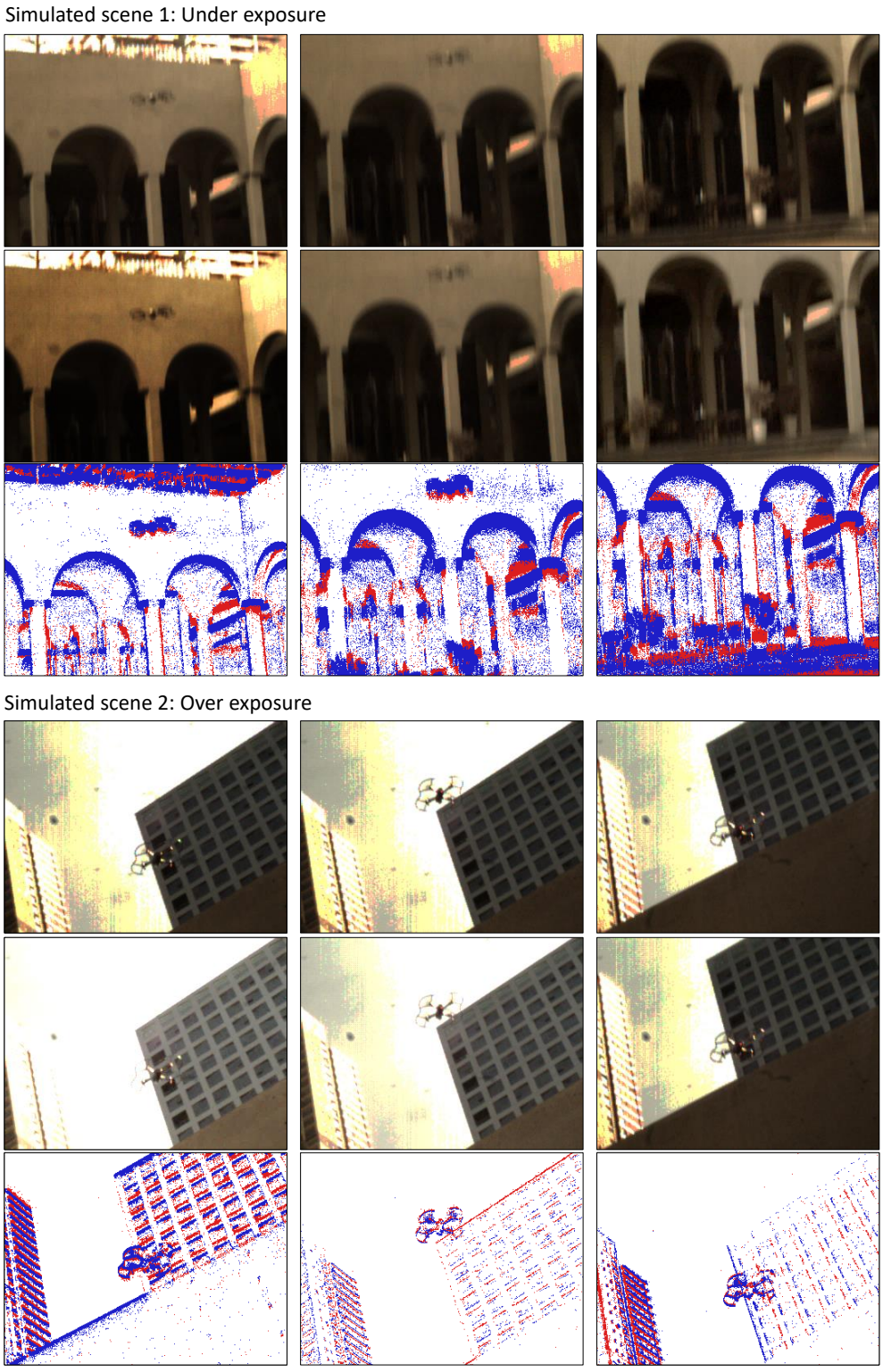

Figure 6. Two exemplar scenes from the simulated high-dynamic-range videos based on the VisEvent dataset. For each scene, we list the original RGB frames, the synthetic high-dynamic-range frames, and the corresponding events from top to bottom. The first scene has a linear increase in intensity and a linear decrease in contrast to mimic underexposure in the 1st frame. The second sequence undergoes linearly decreasing intensity to mimic the overexposure in the first frame.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

Baseline algorithms. We choose 3 different algorithms with state-of-the-art tracking performance as baselines of the proposed solution, i.e., RT-MDNet[45], Siamrpn++[47], and VisEvent[46]. RT-MDNet[45] and Siamrpn++[47] are two RGB-input trackers performing well under normal illumination, and we compare them to verify the performance gains from the event sequence. VisEvent[46] constructs a two-modality neural network fusing RGB and event signals, and we compare with it to verify the effectiveness of the image enhancement module under harsh illumination.

Training. The training is implemented on the NVIDIA 3090 for about 4.7 hours. We set the input image size as well as the spatial resolution of the event sequence to 640×480 pixels and 7 continuous RGB frames (∼350ms) for intensity balancing. We use the Adam optimizer with the learning rate being $5e^{-4}$, the momentum being 0.9, and the weight decay being $5e^{-4}$.

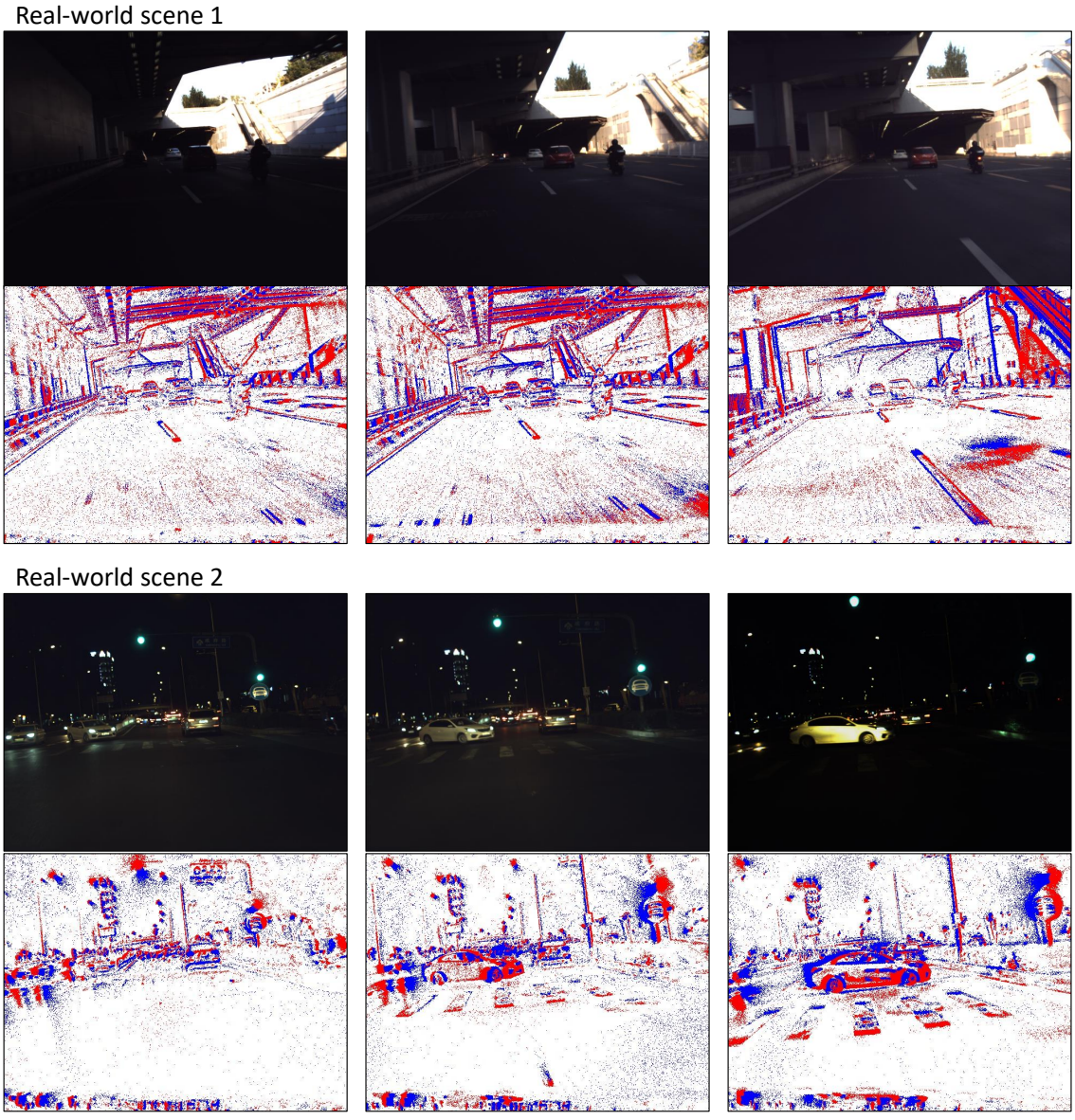

Figure 7. Two typical examples from the real-world dataset captured in harshly lit traffic scenarios. The first scene is captured under a bridge during the daytime, while the second scene is collected in the night at a crossroad. For each scene, the RGB and event cameras are pre-calibrated for pixel-wise registration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

3.2. Results on the simulated data

In this subsection, we validate our approach in terms of image enhancement and object tracking accuracy, based on the simulated data. Here we give both qualitative and quantitative experimental results to comprehensively analyze the effectiveness of the proposed solution. In the qualitative results, we show the result of image enhancement first and compare the object tracking performance to the baseline algorithms later. In the quantitative results, we compare the Precision Plot(PP) and Success Plot(SP) to assess the tracking performance.

3.2.1. Qualitative results

Figure 8 shows the qualitative results on an exemplar video from the simulated dataset. The top row shows the raw RGB sequence, with large intensity changes both within and across frames. In this scene, a person runs from a location with strong illumination toward a destination with a large shadow. Due to the extremely dark intensity, it is challenging to recognize his silhouette in the last frame. We enhance the RGB frames according to the temporal matching extracted from the event signals, and the results are shown in the middle row. The enhanced version is of much more balanced intensity and can highlight the human profile even under weak illumination.

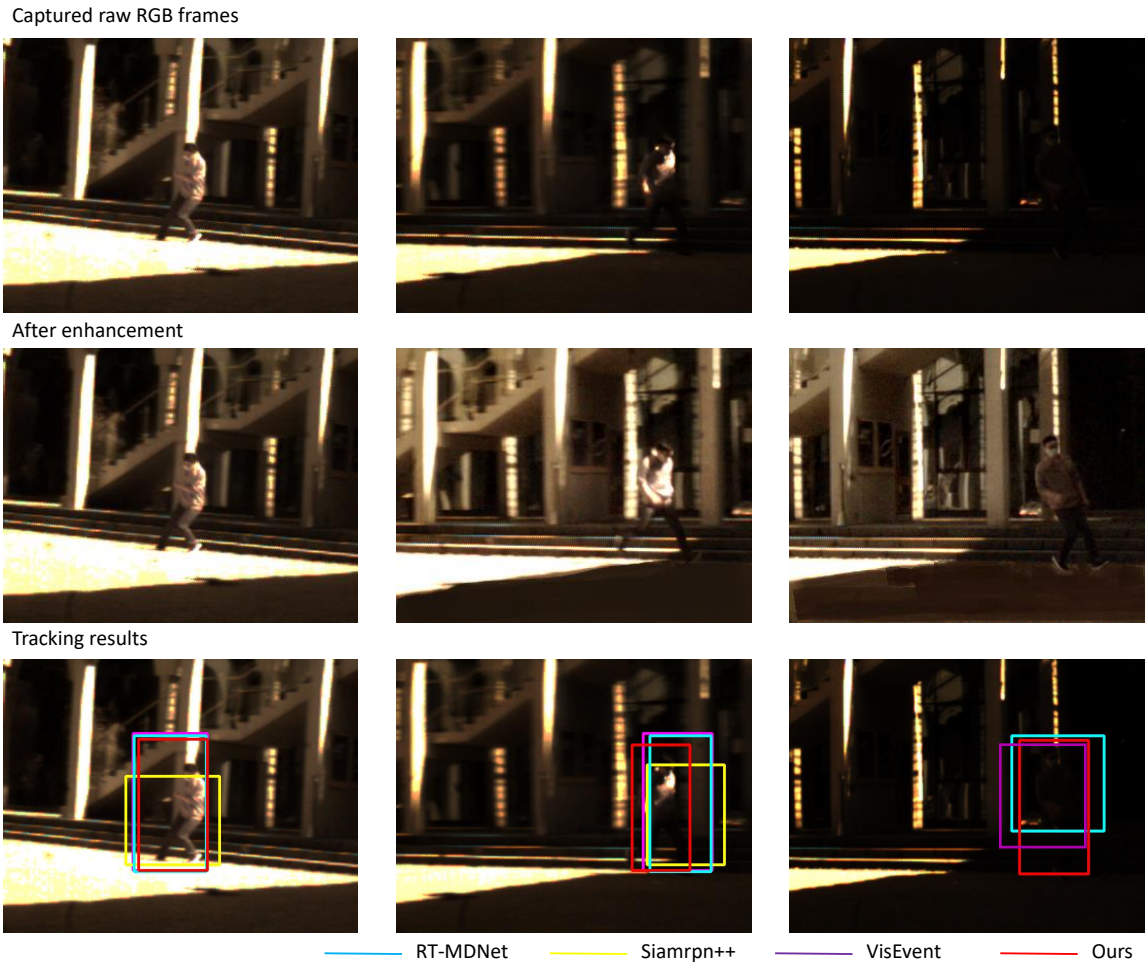

Figure 8. Performance of RGB enhancement of object tracking on a typical exemplar scene in the simulated dataset. Top: The captured RGB video frames. Middle: The corresponding enhanced images by the proposed method. Bottom: The tracking results of different object tracking algorithms.

We further show the object tracking result in the bottom row. The bounding boxes of our approach and the other three competitors are overlaid, with different colors. When sufficiently illuminated, all

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

the algorithms can track the object at high accuracy. RT-MDNet, VisEvent, and the proposed algorithm are comparable, while there exists some deviation in the bounding box output by Siamrpn++ tracks. When the light becomes weak, the proposed algorithm can still identify the person's location, though RT-MDNet and VisEvent's bounding boxes have deviated. When the light is extremely weak, only the proposed method, RT-MDNet, and VisEvent can track the object, because of the high sensitivity and robustness to abrupt intensity changes of the event signals. In comparison, the RGB image in RT-MDNet and VisEvent is not enhanced and thus reduces the final tracking accuracy, while our approach demonstrates reliable tracking consistently.

3.2.2. Quantitative results

We introduce the typical matrix PP and SP here to evaluate the accuracy of object tracking. Specifically, PP indicates the frame percentage where the deviation between the estimated object center location and ground truth is less than the determined threshold. SP denotes the frame percentage that the IoU between the estimated bounding box and the ground truth bounding boxes is higher than the determined threshold. Table 1 shows the PP and SP of our approach and three state-of-the-art object tracking algorithms.

Table 1. The quantitative performance of different object tracking algorithms on the simulated dataset, in terms of PP and SP.

	Our algorithm	VisEvent	Siamrpn++	RT-MDNet
PP	**0.783**	0.712	0.390	0.405
SP	**0.554**	0.465	0.232	0.321

According to Table 1, the proposed algorithm demonstrates the optimal tracking results. Even under harsh illumination, we can track the target object continuously, while Siamrpn++ and RT-MDNet show poor tracking results under the same conditions. Moreover, though VisEvent takes the event signal as the input, it ignores the influence of the low-quality RGB frames and produces inferior tracking accuracy. From the ranking, we can draw two conclusions: first, the event signals can help address the performance degeneration in high dynamic range scenes; secondly, enhancing the degraded RGB frames can further raise the accuracy of object tracking.

3.3. Results on the real-world data

To investigate the performance of our approach in real high-dynamic-range scenes, we test our algorithm on some videos under challenging illumination, with one typical example shown in Figure 9. The video is captured at the tunnel entrance, and the frames in the top row show a car traveling through the tunnel. When the car enters the tunnel, it is difficult to capture images with high visual quality due to insufficient light, and the car turns indistinguishable in the last frame. The middle row shows the result of image enhancement, demonstrating that the visual quality of the RGB frames is largely raised than the raw input.

The tracking results are shown in the bottom row of Figure 9. All four algorithms can track the car at high brightness. When the light becomes weaker, the performance of the two RGB-based tracking algorithms decreases—Siamrpn++ cannot track the car and RT-MDNet produces a bounding box with a large offset, while VisEvent can achieve relatively higher robustness but the bounding box is not accurate. On the contrary, we can achieve reliable tracking over the whole sequence. Based on the above experiments, we can further verify that i) the illumination condition affects the accuracy of object tracking and ii) the event signal can assist the object tracking under harsh illumination.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

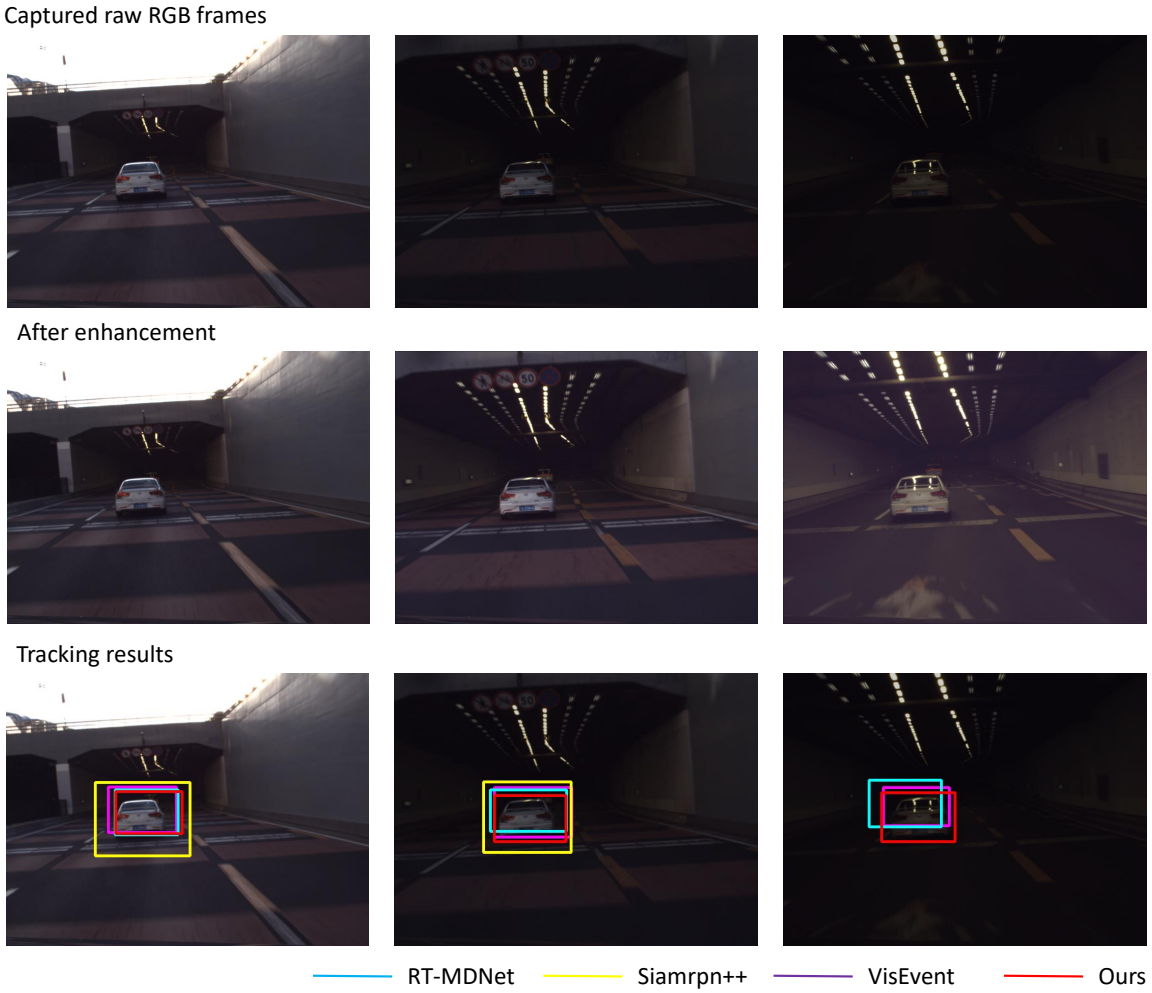

Figure 9. Demonstration of our image enhancement of tracking result, and performance comparison with existing object tracking algorithms on a real-world high-dynamic-range scene—a white car driving through a tunnel. Top: the captured RGB frames. Middle: our enhanced RGB images. Bottom: the tracking results of different algorithms.

3.4. Ablation studies

The ablation experiment focuses on validating the contribution of event-based temporal alignment to RGB image enhancement and object tracking. In the proposed approach, we use Harris corner detection to retrieve key feature points from the dense event sequence, and here we compared its performance against two methods—using random event signals as key features and the detected Harris corner points from the RGB images rather than event signals.

From the upper row in Figure 10, one can see that there exist large intensity variations within each frame and abrupt changes among frames, which is quite challenging for object tracking algorithms and even human vision systems, especially in the third frame. Here we adopt the person in the third frame as the tracking target, and the results with different key feature guidance are shown at the bottom row. One can see that the proposed alignment strategy performs best in terms of both the quality of the enhanced image and object tracking accuracy. In comparison, the result produced by registration from random event signals slightly enhances the image quality and results in a looser bounding box while registration from RGB frames provides little help, which again validates the strategy of introducing event cameras for such harshly lit scenes. The inferior performance of the two benchmarking implementations is mainly attributed to the fact that they cannot identify the temporal matching properly due to the lack of descriptive features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2023 doi:10.20944/preprints202312.1056.v1

Figure 10. An example showing the results of ablation studies. The upper row displays the RGB frames of a high-dynamic-range scene. The lower row shows the image enhancement and object tracking results based on three different temporal registration guidance, with the person in the third frame (darkest and most challenging) as the target object. From left to right: key feature alignment from the proposed event-based Harris corner points, random event signals, and Harris corner points in RGB frames.

4. Summary and Discussions

Visible-wavelength optical cameras provide rich scene information for the environmental sensing of drones. However, the harsh illumination causes high dynamic range (e.g., at nighttime, at entrance or exit, etc.) and hampers reliable environmental perception. In order to extend the applicability of visible-wavelength cameras in real scenes, we propose a dual-sensing architecture that leverages the advantages of event cameras to raise the imaging quality of the RGB sensor as well as the successive object tracking performance.

The proposed event-assisted robust object tracker exploits two main features of event signals, i.e., robust imaging under complex illumination and fast response. These advantageous and unique features support extracting the continuous trajectories of corner points to guide the temporal registration of high-dynamic-range RGB frames. The registration plays a central role in compensating the intensity changes. Experimentally, the proposed event-assisted robust object tracking can work quite well in a high-dynamic-range environment that goes beyond the capability of RGB cameras.

The performance of the proposed algorithm is superior to both the counterparts taking only the RGB frames as input or directly taking two inputs, and the advantages hold in a wide range of applications. From the comparison, we can get the following two conclusions: (i) Under harsh illumination, the quality of RGB images greatly affects the performance of the downstream tasks. In order to ensure the robustness of the performance of tasks such as object tracking, the RGB frames need to be enhanced first. (ii) Event signals, as a lightweight and efficient sensor, can be used to capture critical information in high-speed moving scenes. In addition, event signals are insensitive to lighting conditions and can be used for scene sensing under extreme illumination.

In the future, we will dig deeper into the characteristics of event signals and construct neural networks that are more compatible with event signals to realize lightweight network design and

efficient learning. In addition, we will integrate sensing units such as LIDAR and IMU to achieve depth-aware 3D representation of the scene.

Author Contributions: Y. H. and J. S. conceived this project. Y. H. designed the framework and the network architecture. Y. H. and X. Y. implemented the event-based key feature alignment as well as temporal RGB image enhancement. H. L. collected the dataset and conducted the comparison experiments. X. Y designed and conducted the ablation studies and analyzed the experiment results. J. S. dominated the discussion of this work. J. S. supervised this research and finally approved the version to be submitted. All the authors participated in the writing of this paper.

Funding: This work is jointly supported by the National Natural Science Foundation of China [grant numbers 61931012].

Data Availability Statement: The data are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, J.; Hu, J.; Lian, J.; Fan, Z.; Ouyang, X.; Ye, W. Seeing the forest from drones: Testing the potential of lightweight drones as a tool for long-term forest monitoring. *Biological Conservation* **2016**, *198*, 60–69.
2. Duffy, J.P.; Cunliffe, A.M.; DeBell, L.; Sandbrook, C.; Wich, S.A.; Shutler, J.D.; Myers-Smith, I.H.; Varela, M.R.; Anderson, K. Location, location, location: considerations when using lightweight drones in challenging environments. *Remote Sensing in Ecology and Conservation* **2018**, *4*, 7–19.
3. Zhang, Y.; He, D.; Li, L.; Chen, B. A lightweight authentication and key agreement scheme for Internet of Drones. *Computer Communications* **2020**, *154*, 455–464.
4. McNeal, G.S. Drones and the future of aerial surveillance. *The George Washington Law Review* **2016**, *84*, 354.
5. Akram, M.W.; Bashir, A.K.; Shamshad, S.; Saleem, M.A.; AlZubi, A.A.; Chaudhry, S.A.; Alzahrani, B.A.; Zikria, Y.B. A secure and lightweight drones-access protocol for smart city surveillance. *IEEE Transactions on Intelligent Transportation Systems* **2021**, *23*, 19634–19643.
6. Guvenc, I.; Koohifar, F.; Singh, S.; Sichitiu, M.L.; Matolak, D. Detection, tracking, and interdiction for amateur drones. *IEEE Communications Magazine* **2018**, *56*, 75–81.
7. Bamburry, D. Drones: Designed for product delivery. *Design Management Review* **2015**, *26*, 40–48.
8. Panda, S.S.; Rao, M.N.; Thenkabail, P.S.; Fitzerald, J.E. Remote Sensing Systems—Platforms and Sensors: Aerial, Satellite, UAV, Optical, Radar, and LiDAR. In *Remotely Sensed Data Characterization, Classification, and Accuracies*; CRC Press, 2015; pp. 37–92.
9. Jeong, N.; Hwang, H.; Matson, E.T. Evaluation of low-cost lidar sensor for application in indoor UAV navigation. In Proceedings of the IEEE Sensors Applications Symposium. IEEE, 2018, pp. 1–5.
10. Bellon-Maurel, V.; McBratney, A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils–Critical review and research perspectives. *Soil Biology and Biochemistry* **2011**, *43*, 1398–1410.
11. Chen, P.; Dang, Y.; Liang, R.; Zhu, W.; He, X. Real-time object tracking on a drone with multi-inertial sensing data. *IEEE Transactions on Intelligent Transportation Systems* **2017**, *19*, 131–139.
12. Wen, L.; Zhu, P.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Liu, C.; Cheng, H.; Liu, X.; Ma, W.; et al. Visdrone-SOT2018: The vision meets drone single-object tracking challenge results. In Proceedings of the European Conference on Computer Vision Workshops, 2018, pp. 0–0.
13. Bartak, R.; Vykovskỳ, A. Any object tracking and following by a flying drone. In Proceedings of the Mexican International Conference on Artificial Intelligence. IEEE, 2015, pp. 35–41.
14. Zhang, H.; Wang, G.; Lei, Z.; Hwang, J.N. Eye in the sky: Drone-based object tracking and 3D localization. In Proceedings of the ACM International Conference on Multimedia, 2019, pp. 899–907.
15. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision Workshops. Springer, 2016, pp. 850–865.
16. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. ECO: Efficient convolution operators for tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 6638–6646.

17. Dai, K.; Wang, D.; Lu, H.; Sun, C.; Li, J. Visual tracking via adaptive spatially-regularized correlation filters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4670–4679.
18. Li, P.; Chen, B.; Ouyang, W.; Wang, D.; Yang, X.; Lu, H. GradNet: Gradient-guided network for visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6162–6171.
19. Mitrokhin, A.; Fermüller, C.; Parameshwara, C.; Aloimonos, Y. Event-based moving object detection and tracking. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2018, pp. 1–9.
20. Chen, H.; Suter, D.; Wu, Q.; Wang, H. End-to-end learning of object motion estimation from retinal events for event-based object tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 10534–10541.
21. Debevec, P.E.; Malik, J. Recovering high dynamic range radiance maps from photographs. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*; 2023; pp. 643–652.
22. Kang, S.B.; Uyttendaele, M.; Winder, S.; Szeliski, R. High dynamic range video. *ACM Transactions on Graphics* **2003**, *22*, 319–325.
23. Mangiat, S.; Gibson, J. High dynamic range video with ghost removal. In Proceedings of the Applications of Digital Image Processing. SPIE, 2010, Vol. 7798, pp. 307–314.
24. Mangiat, S.; Gibson, J. Spatially adaptive filtering for registration artifact removal in HDR video. In Proceedings of the IEEE International Conference on Image Processing. IEEE, 2011, pp. 1317–1320.
25. Kalantari, N.K.; Ramamoorthi, R.; et al. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics* **2017**, *36*, 144–1.
26. Gryaditskaya, Y. High dynamic range imaging: problems of video exposure bracketing, luminance calibration and gloss editing **2016**.
27. Hajisharif, S.; Kronander, J.; Unger, J. Adaptive dualISO HDR reconstruction. *EURASIP Journal on Image and Video Processing* **2015**, *2015*, 1–13.
28. Heide, F.; Steinberger, M.; Tsai, Y.T.; Rouf, M.; Pająk, D.; Reddy, D.; Gallo, O.; Liu, J.; Heidrich, W.; Egiazarian, K.; et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics* **2014**, *33*, 1–13.
29. Cai, J.; Gu, S.; Zhang, L. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing* **2018**, *27*, 2049–2062.
30. Nayar, S.K.; Mitsunaga, T. High dynamic range imaging: Spatially varying pixel exposures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2000, Vol. 1, pp. 472–479.
31. Zhao, H.; Shi, B.; Fernandez-Cull, C.; Yeung, S.K.; Raskar, R. Unbounded high dynamic range photography using a modulo camera. In Proceedings of the IEEE International Conference on Computational Photography. IEEE, 2015, pp. 1–10.
32. Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.J.; Conradt, J.; Daniilidis, K.; et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *44*, 154–180.
33. Muglikar, M.; Gehrig, M.; Gehrig, D.; Scaramuzza, D. How to calibrate your event camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1403–1409.
34. Lagorce, X.; Meyer, C.; Ieng, S.H.; Filliat, D.; Benosman, R. Asynchronous event-based multikernel algorithm for high-speed visual features tracking. *IEEE Transactions on Neural Networks and Learning Systems* **2014**, *26*, 1710–1720.
35. Rebecq, H.; Ranftl, R.; Koltun, V.; Scaramuzza, D. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *43*, 1964–1980.
36. Brandli, C.; Muller, L.; Delbruck, T. Real-time, high-speed video decompression using a frame-and event-based DAVIS sensor. In Proceedings of the IEEE International Symposium on Circuits and Systems. IEEE, 2014, pp. 686–689.
37. Ni, Z.; Pacoret, C.; Benosman, R.; Ieng, S.; RÉGNIER*, S. Asynchronous event-based high speed vision for microparticle tracking. *Journal of Microscopy* **2012**, *245*, 236–244.

38. Tulyakov, S.; Gehrig, D.; Georgoulis, S.; Erbach, J.; Gehrig, M.; Li, Y.; Scaramuzza, D. Time lens: Event-based video frame interpolation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16155–16164.
39. Tulyakov, S.; Bochicchio, A.; Gehrig, D.; Georgoulis, S.; Li, Y.; Scaramuzza, D. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17755–17764.
40. Pan, L.; Liu, M.; Hartley, R. Single image optical flow estimation with an event camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020, pp. 1669–1678.
41. Bardow, P.; Davison, A.J.; Leutenegger, S. Simultaneous optical flow and intensity estimation from an event camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 884–892.
42. Wan, Z.; Dai, Y.; Mao, Y. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing* **2022**, *31*, 7237–7251.
43. Akolkar, H.; Ieng, S.H.; Benosman, R. Real-time high speed motion prediction using fast aperture-robust event-driven visual flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *44*, 361–372.
44. Ramesh, B.; Zhang, S.; Lee, Z.W.; Gao, Z.; Orchard, G.; Xiang, C. Long-term object tracking with a moving event camera. In Proceedings of the The British Machine Vision Conference, 2018, p. 241.
45. Jung, I.; Son, J.; Baek, M.; Han, B. Real-time MDNet. In Proceedings of the European Conference on Computer Vision, 2018, pp. 83–98.
46. Wang, X.; Li, J.; Zhu, L.; Zhang, Z.; Chen, Z.; Li, X.; Wang, Y.; Tian, Y.; Wu, F. VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows. *IEEE Transactions on Cybernetics* **2023**, pp. 1–14.
47. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4277–4286.